Phylogenomic and Structure-Function Relationship Studies of Proteins Involved in EBV Associated Oncogenesis

by

Mushtaq Hussain

The research presented in this thesis was sponsored by Dow University of Health Sciences, Karachi, Pakistan and Higher Education Commission, Pakistan and is author's original work unless otherwise stated.

Mushtaq Hussain

In the memory of my Paternal Grand Father and Maternal Grand Mother

Wali Bhai Dadu Bhai Junagadh Wala (Late)

Hawan Ji Bai Chand Ibrahim (Late)

Abstract

This study covers the investigation of evolutionary and structure-function relationship aspects of several cancer related proteins. One part of the study deals with the investigation of a critical protein of Epstein-Barr Virus (EBV) the Nuclear Antigen 1 (EBNA1), and its interactions with different host proteins. One of these host proteins is a member of a large gene family, encoding ubiquitin specific proteases (USP), known as USP7. The second section of the thesis deals with the molecular evolution of the USP gene family. Another set of cellular proteins deregulated during EBV associated oncogenesis are members of the glycoside hydrolase (GH18) family. Their phylogenetic relationships and protein structures were investigated in the third section of this thesis.

EBNA1 is the only EBV protein that consistently expressed in all latent forms of the EBV infections. The protein is involved in the genome maintenance and a substantial body of evidence suggests that it has a role in EBV associated oncogenesis. In this study, full length molecular models of the EBNA1 protein were generated using the programmes, I-TASSER, MOE and Modeller. The best models were selected on the basis of plausibility in structural and thermodynamical parameters and from this models of EBNA1 homologues of primates lymphocryptoviruses (LCVs) were generated. The C-terminal DNA binding and homodimerisation domain was predicted to be structurally similar between different LCV EBNA1 homologues, indicative of functional conservation. The central glycine alanine repeat (GAr) domain was predicted to be primarily composed of α helices, while almost all of the protein interaction region was found to be unstructured, irrespective of the prediction approach used and sequence origin. Predicted USP7 and Casein kinase 2 (CK2) binding sites and GAr were observed in the EBNA1 homologues of Old World primate LCVs, but not in the marmoset homologue suggesting the co-evolution of both these sites. Dimer conformations of the EBNA1 monomer models were constructed using SymmDock, where the C-terminal tail was predicted to wrap around the proline rich loop of another monomer, possibly contributing to dimer stability. This feature could be exploited in therapeutic design, hence an inhibitor peptide was designed and a preliminary evaluation was conducted to explore its ability to inhibit EBNA1 function in cell survival. The peptide array libraries of EBNA1 were used to investigate the binding regions and critical contact points between EBNA1 and partner proteins. Human EBP2 and USP7 proteins were expressed in bacteria and probed on the EBNA1 array. The data confirm the previously known binding region for EBNA1-EBP2 and EBNA1-USP7 interactions. In addition further information was gained regarding the critical contact residues and the potential role of phosphorylation of serine residues of EBNA1 in its binding with EBP2 and USP7.

The human genome encodes nearly 100 USPs which contribute to regulate the turnover of cellular proteins. These homologues are divided into 16 paralogous groups, all sharing a characteristic peptidase C19 domain. Evolutionary relationships between these homologues were explored by datamining and the phylogenetic reconstruction of peptidase C19 domain sequences. The data reveal an ancient relationship between the genes, with expansion occurring throughout the course of evolution, but particularly at the base of the vertebrates, at the time of the two whole genome duplications. A comparison between the phylogenetic architecture and protein interaction networks suggests the parallel emergence of many molecular pathways and the associated USPs.

The GH18 gene family includes chitinases and related non catalytic proteins. Most mammals encode at least three chitinases (CHIT1, CHIA/AMCase and CTBS), as well as several homologues encoding catalytically inactive chitinase-like proteins or chilectins. Phylogenomic analysis shows that the family has undergone extensive expansion, initiating with a duplication event at the root of the vertebrate tree, resulting in the origin of the ancestors of CHIT1 and CHIA. Two further duplications of ancestral CHIA predate the divergence of bony fishes, one leading to a newly identified paralogous group (we have termed CHIO). In tetrapods, additional CHIA duplications predate and postdate the amphibian/mammalian split and relics of some exist as pseudogenes in the human genome. Homology modelling of structurally unresolved GH18 homologues in mouse and human was conducted using Modeller and I-TASSER. All resolved and predicted structures share a TIM barrel $(\beta/\alpha)_8$ and $\alpha+\beta$ domain. A central ligand binding cavity was also found in all GH18 homologues. The variation in size and shape of different paralogous proteins, indicate the difference in their ligands specificity and in turn potential functions.

Acknowledgements

The present work would not be possible without the invaluable support of many individuals. Firstly, I would like to express my deepest gratitude to my supervisor ***Dr Joanna Beatrice Wilson*** for her perpetual extension of ideas, feedbacks and freedom I was given during my studies. She has all the attributes to sculpt a heavily dependent student like me into an independent researcher. Her immediate recognition and understanding of my interests and strengths allows me to switch from a immunology based project to a bioinformatics based research. Her artistic skills, as evident from her paintings and interior decoration of her house and garden, came very handy in beautifying many of the figures in this thesis and associated publications. It would be impossible to imagine the shape of the present thesis without her extensive help in editing, proofreading and more importantly filtering out the scientific errors. In short, it will be my honour to take my experience with her, as a role model for supervision in the impending future. Her thanks could not be complete without mentioning the regular supply of homemade chilly pickle (from home grown chillies), lemon tea (when I catch a cold) and coffee during my late stays in the lab. And of course, trips to Loch Lomond and “A play A pie and A pint” at Oran Mor with her are some of the most cherishable memories of mine. Finally, my experience with her as a demonstrator provided me an opportunity to learn skills to engage students especially with reference to critical thinking.

My sincere thanks are due to ***Prof. Dr. George Baillie***, ***Ruth MacLeod*** and ***Ryan Cameron*** for manufacturing the peptide arrays and providing the expression vectors. I am also grateful to ***Dr. Mark Bailey*** for his ever present help in conducting the phylogenetic studies. I always found his joyful and constructive criticism immensely helpful in undertaking phylogenomic studies. ***Dr. Derek Gatherer*** is a real help in constructing the molecular models of EBNA1 and in inferring the results pragmatically.

I would like to extend my thanks to ***Donald Campbell***, who is beyond any doubt the most resourceful person when it comes to lab issues. His constant supply of documentaries, movies and music made my weekends informative and entertaining. Without him I would not be able to understand dynamics of World War II and the events revolving around it. The friendship with him is the treasure which I will always relish. I am very thankful to all the lab members of Lab412 especially ***Jin*** (for generating some of the array results) and ***Xiao Gao***. This acknowledgement would not be complete without mentioning, my lab colleague ***Saeed Al-Ghamdi*** (dubbed The King; Ya Marhaba) from the Kingdom of Saudia Arabia. His generous and brotherly help is one of the biggest reason I was able to complete this study. My trip to Loch Tay, Rosslyn Chapel and many other places in UK are only become possible because of him. In short, wherever I need, I had found him amongst the first to support. I cannot miss mentioning my dearest friend ***Dr. Nusrat Jabeen*** for being always remain in touch. Keeping me informed about the folks in Pakistan and independently sorting out many of the formalities for me in Pakistan. To me, the friendship with Nusrat is an example of unconditional sincerity.

My thanks is also due to the ***Higher Education Commission, Govt. of Pakistan*** and ***Dow University of Health Sciences***, ***Pakistan*** for sponsoring the higher studies of mine and many other of my country fellows despite severe economical constraints.

Being first time abroad, it is not easy for me and for my family, I had left behind in Pakistan. I am indebted to my family members, ***Muhammad Hussain*** (Dad), ***Mumtaz Hussain*** (Mom), ***Dr. Saba Hussain*** (Sis) and ***Zohaib Hussain*** (Bro) for taking care of my wife and kids and enduring my absence from the home.

No words of thanks could truly reflect the unconditional love and support I have gained from my better half, wife, ***Sofia Seher Hussain*** (Sofo). Her patience for understanding my long absence from home and fulfilling both of our duties for taking care of our kids is inconceivable for me. Finally, my thanks to the ultimate moral boosters ***Zaeem Hussain*** (Bundo Balay; son) and ***Umaima Hussain*** (Taman Rani; daughter) for being so naughty yet so smart. Guys you are all my reasons.

1. Dr. Joanna B. Wilson
2. Dr. Mark Bailey
3. Donald Campbell
4. Prof. George Bailey
5. Dr. Derek Gatherer
6. Zaeem Hussain
7. Sofia Seher Hussain
8. Nusrat Jabeen
9. Umaima Hussain
10. Muhammad Hussain
11. Mumtaz Hussain
12. Saeed Alghamdi
13. Jin

Table of Contents

C:\Users\ABM\Desktop\S&H\PhD Research\Thesis\PhD Thesis4.docx - _Toc378456924

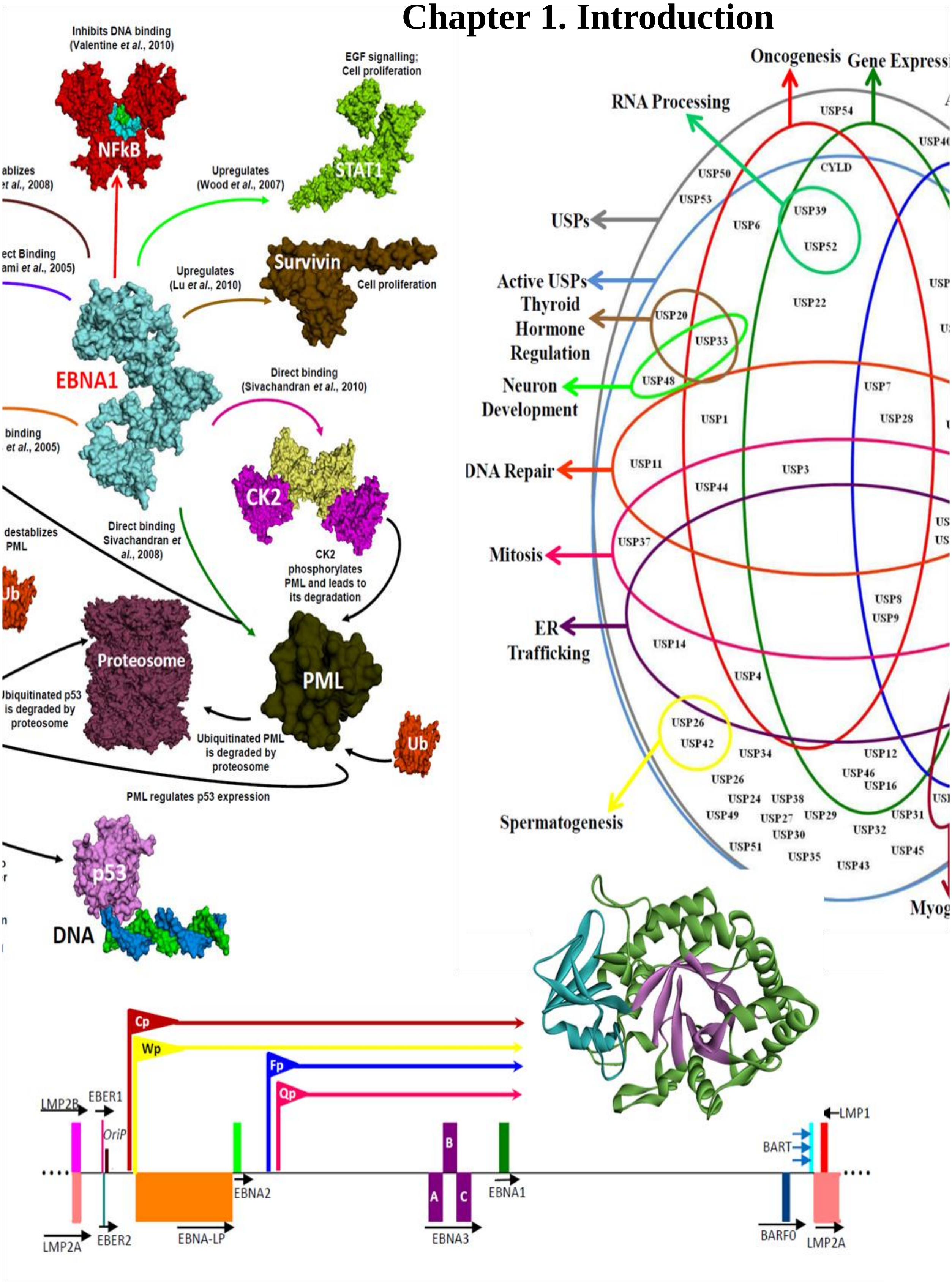
List of Figures

List of Tables

Abbreviations

~	Approximately
α	Alpha
β	Beta
µl	Micro litre
Å	Degree Angstrom
^{o}C	Degree Celsius
µM	Micro molar
APS	Ammonium per sulphate
BLB	B lymphoblasts
bp	Base pairs
BSA	Bovine Serum Albumin
BDCA4DC	BDCA4 dendritic cells
CD	Cluster of differentiation
CD4+T	Cluster of differentiation 4+ T cells
CD8+T	Cluster of differentiation 8+ T cells
CD19B	Cluster of differentiation 19+ B cells
CD56+NK	Cluster of differentiation 56+ natural killer cells
CD105+EC	Cluster of differentiation 105+ endothelial cells
CD71+ EE	Cluster of differentiation 71+ early erythroid
CK2α	Casein kinase 2 alpha
CK2β	Casein kinase 2 beta
DAPI	4',6-diamidino-2-phenylindole
DEPC	Diethylpyrocarbonate
dH$_2$O	Distilled water
dNTP	Deoxynucleotide phosphate
DTT	dithiothreitol
DUB	Deubiquitinating enzymes
EBV	Epstein-Barr Virus
EBNA1	Epstein-Barr Virus Nuclear Antigen 1
EBP2	Epstein-Barr Virus Nuclear Antigen 1 Binding protein 2
EDTA	Ethylene diamine tetra acetic acid
FBS	Fetal Bovine Serum
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LB	Lysogeny broth
LCL	Lymphoblastoid cell line
M	Molar
mg	Milligram
miRNA	Micro RNA
ml	Millilitre
mM	Millimolar
MOPS	3-(N-Morpholino)propanesulfonicacid,4-Morpholinepropanesulfonic acid
NFM	Non fat milk
PBS	Phosphate buffer saline
PBST	Phosphate buffer saline with tween 20
PCR	Polymerase chain reaction
SDS	Sodium do-Decyl Sulphate
SUMO	Small Ubiquitin-like Modifier
SDS-PAGE	Sodium do-Decyl Sulphate Polyacrylamide Gel Electrophoresis
TAE	Tris acetic acid EDTA buffer
TBS	Tris phosphate buffer saline
TBST	Tris buffer saline with tween 20
TEMED	N,N,N’,N’-tetramethylethylenediamine
USP	Ubiquitin Specific Peptidase/Protease

Chapter 1. Introduction



1.Introduction

The work in this thesis covers the investigation of evolutionary and structure-function aspects of several cancer related proteins. One part of the study deals with a critical protein of Epstein-Barr Virus (EBV), the EBV Nuclear Antigen 1 (EBNA1). An important cellular protein that interacts with EBNA1 is the ubiquitin specific protease 7 (USP7), the second section of the thesis explores the phylogenomics of the USP gene family. Another set of cellular proteins deregulated during EBV associated oncogenesis are members of the glycoside hydrolase-18 (GH18) family and the third section of this thesis investigated GH18 gene family phylogenomics and protein structure.

1.1. Epstein-Barr Virus

Epstein-Barr Virus (EBV), also called human herpesvirus 4 (HHV-4), is a linear double stranded DNA containing virus. Taxonomically, it belongs to genus *lymphocryptovirus* (LCV) or gamma-1 herpesvirus of family *herpesviradae* and subfamily *gammaherpesvirinae*. The virus normally infects B cells and can undergo a lytic cycle, leading to the release of viral particles, and a latent cycle in which viral genome is stably maintained within the infected cells. The genome is around 172Kbp in size and contains approximately 100 distinct genes (Baer *et al*., 1984; Kieff and Rickinson, 2007). Out of these, 11 genes are expressed early in viral infection and variably in different latent states. These include: six Epstein-Barr virus Nuclear Antigens: *EBNA1*, *2*, *3A*, *3B*, *3C* and *–LP*, three Latent Membrane Proteins: *LMP1*, *LMP2A* and *LMP2B,* two Epstein-Barr virus Encoded RNAs: *EBER1* and *EBER2* as well as multiple microRNAs. The viral genome is surrounded by a protein capsid and between the capsid and inner envelope lies a protein tegument which anchors several glycoprotein that define cell tropism, host range and receptor recognition of the virus (Kieff and Rickinson, 2007). To date, two viral subtypes, type 1 and type 2 have been identified which differ at EBNA loci. The two types also differ in their transforming ability (Takimoto *et al*., 1989) and epidemiology (Kieff and Rickinson, 2007) as EBV type 1 is prevalent in many parts of the world whereas type 2 is more prevalent in Africa. Since the global infection rate of EBV is more than 90%, it is included among the most successful viruses in evolutionary terms (Cohen, 2000). EBV was the first virus proposed to be associated with the human cancers (Epstein *et al*., 1964) but absolute recognition of the oncogenic potential of EBV and its association with other human disorders is still a growing area.

1.2. Brief time line of EBV research

EBV was first discovered by Epstein and co-workers using electron microscopy of a cell line derived from a Burkitt’s lymphoma patient (Epstein *et al*., 1964). Since its discovery, a wealth of information has been gathered that enable us to understand many key aspects of EBV biology. One reflection of this is the presence of over 30,000 research articles (papers and reviews) in the National Center for Biotechnology Information (NCBI) database having the key word Epstein-Barr Virus (Figure 1.1).

1.3. EBV infections in humans

Normally, EBV infections are asymptomatic and the virus is spread via saliva. The primary site of infection is the oropharynx where the virus comes into contact with B lymphocytes in the lymphoid tissue of Waldeyer's ring. However, it is now increasingly evident that EBV can also infect epithelial cells, T cells, natural killer (NK) cells, smooth cells and monocytes (reviewed in Hutt Fletcher, 2007).

Initial attachment of EBV is predominantly mediated by the interaction between its envelope protein (gp350/220) and the cellular complement component receptor 2 (CR2/CD21) a protein found on the B cell surface (Fingeroth *et al*., 1984; Johannsen *et al*., 2004). However, a gp350 deletion mutant of EBV retains the ability to transform B cells, although with much reduced efficiency, suggesting other portal(s) may also present to facilitate EBV infection. Nevertheless, gp350 is a major requirement as antibodies to gp350 neutralize the infection of B cells by impairing virus attachment (Tanner *et al*., 1988). Additionally, the structure of CR2 has been resolved by crystallography and critical regions for gp350 binding have been identified (Prota *et al*., 2002). Intriguingly, epithelial cells lack CR2 or express it at a very low level and the underlying mechanism of virus attachment with the epithelial cells is still unclear. However, possible mechanisms include viral attachment via gp350 antibodies binding to the IgA receptor on epithelial cells (Sixbey *et al*., 1992), and attachment of viral glycoproteins gH and gL to an unknown receptor on epithelial cells (Molesworth *et al*., 2000) and binding between the viral membrane protein BMRF2 with integrins on epithelial cells (Tugizov *et al*., 2003).

Virus fusion with either B cells or epithelial cells requires three glycoproteins gH, gL and gB. Briefly, the attachment of gp350/220 with the CR2 receptor potentially triggers signalling events that initiate the process of endocytosis. CR2 switches its binding from gp350 to gp220 which in turn allows gp42 to interact with HLA class II (HLAII). This interaction facilitates the core fusion machinery (gH, gL and gB) to interact with the endosomal membrane allowing cellular internalization of the virus (Hutt-Fletcher, 2007). Viral fusion in epithelial cells is proposed to be independent of the gp42-HLAII interaction but mediated predominantly by the gH, gL and gB complex (Wang *et al*., 1998).

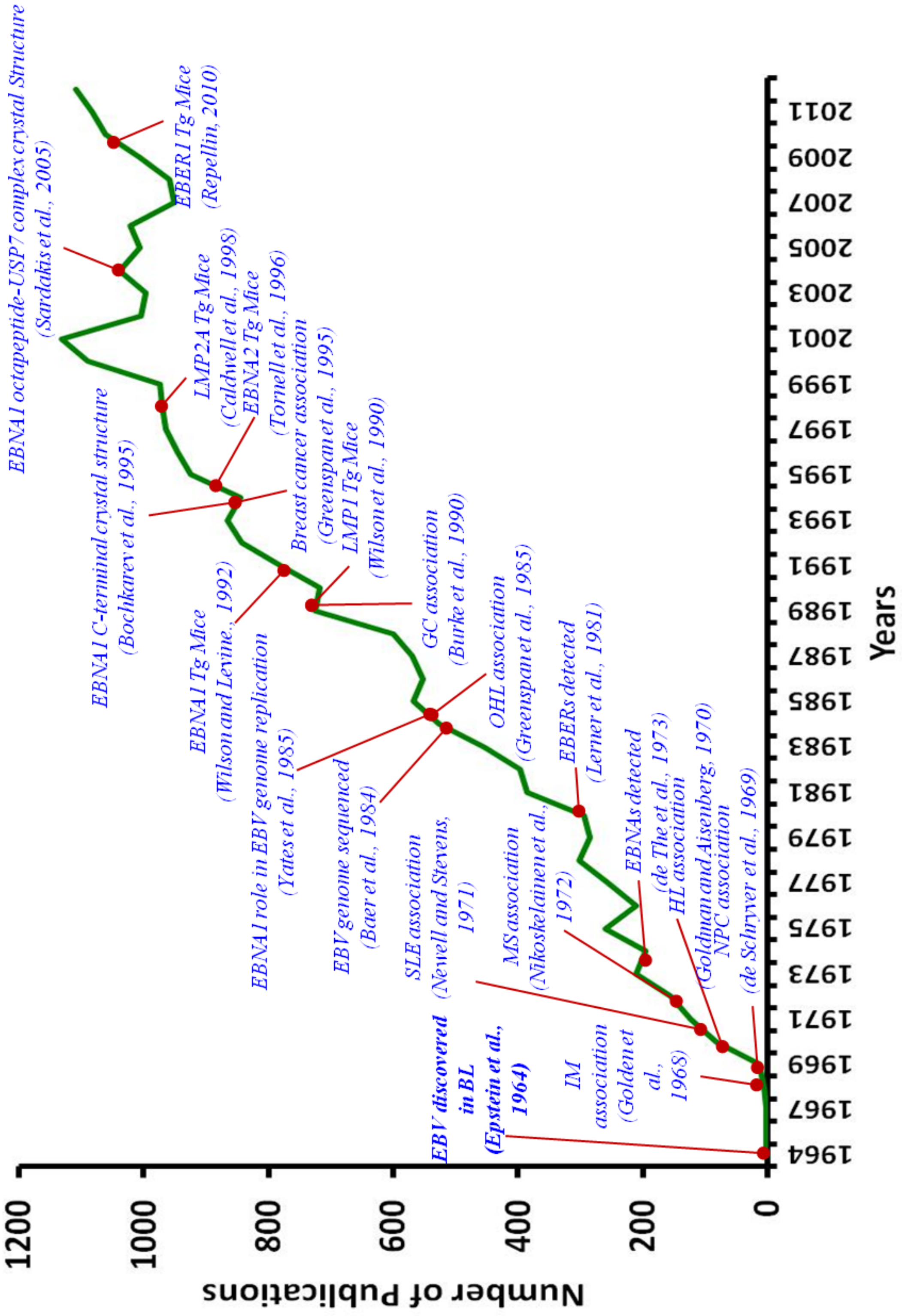


Figure 1.1. Brief timeline of EBV research. The graph demonstrates the continuous increase in the number of scientific publications related to EBV over the years. Note, only the NCBI database was consulted in developing the graph and only some of the key observations are indicated here. Key: Burkitt's lymphoma (BL), Hodgkin's lymphoma (HL), multiple sclerosis (MS), systematic lupus erythematous (SL), oral hairy leukoplakia (OHL), gastric carcinoma (GC), infectious mononucleosis (IM), nasopharyngeal carcinoma (NPC), Epstein Barr Virus Nuclear Antigen (EBNA), Latent Membrane Protein (LMP), Epstein Barr Virus encoded RNA (EBER), Ubiquitin Specific Protease (USP), Transgenic (Tg). Note: the association of EBV with breast cancer and any functional role in this disease is still controversial.

Once inside the cell, EBV can undergo two routes in its life cycle: 1) Lytic infection which is marked by the active propagation of virus in the host. 2) Latent infection which ensures the persistence of the virus in the host without viral particle propagation. The virus can be reactivated into lytic infection from latency.

1.3.1. Lytic infection

After cellular internalization the nucleocapsid is dissolved and the genome enters into the cell nucleus. The lytic cycle is differentiated into three stages: Immediate-early (IE), early (E) and late (L). The IE stage is marked by the expression of *BZLF1* and *BRLF1* genes, the encoded proteins subsequently act as transactivators for other lytic genes and lead to the expression of early stage genes (*BMRF1*, *BALF2*, *BAL55*, *BBL2/3*, *BBLF4* and *BSF1* etc) and subsequently late stage genes for viral structural proteins (*gp350/220*, *VCA*, *gp85*, *gp25* and *gp42*). In this productive life cycle, the EBV genome is amplified by 100 to 1000 fold (Hammerschmidt and Sugden, 1988) and multiple rounds of DNA replication originating at two sites make lytic infection distinct from latent infection (reviewed in Tsurumi *et al*., 2005).

1.3.2. Latent infection

Latent infection of EBV does not support the active propagation of virus but it ensures the stable persistence of the viral genome in the host cell. Unlike lytic infection, replication of the viral genome in latent infection occurs via host DNA polymerase (Amon *et al*., 2005) and from a separate origin, *OriP*. During latent infection the viral genome exists as a closed circular extrachromosomal plasmid or episome, packaged around host histone molecules (Dyson and Farell, 1985) and it stably replicates once during the cell cycle along with the host genome (Kirchmaier and Sugden, 1995). It is interesting to note that all primary or lytic infections of EBV begin with the expression of all latent genes which drives the infected B cell into proliferation. However, soon the expression of these genes is suppressed to evade immune recognition and if entering the lytic cycle, superseded by lytic genes. However, if the virus enters latency, all viral protein shuts off, ultimately entering latency 0. To date four main latency programmes have been categorized on the basis of latent gene expression profile of EBV infected cell lines and in the healthy host: latency 0, I, II and III. Table 1.1. shows latent gene expression pattern in different latency programmes of EBV. Moreover, these are broad categorization and different patterns can be found.

1.4. Spectrum of EBV associated human diseases

In most cases, primary infection of pre-adolescents of EBV lacks any clinical manifestation and is countered by the host immune response. However, the immune system

Latent genes	Latency 0	Latency I	Latency II	Latency III
EBNA1	-	+	+	+
EBNA2	-	-	-	+
EBNA3A	-	-	-	+
EBNA3B	-	-	-	+
EBNA3C	-	-	-	+
EBNA-LP	-	-	+	+
LMP1	-	-	+	+
LMP2A	-	-	+	+
LMP2B	-	-	+	+
EBER1	-	+	+	+
EBER2	-	+	+	+
Diseases	Healthy individuals	Burkitt's lymphoma	Nasopharyngeal carcinoma Hodgkin's lymphoma	Infectious mononucleosis, PTLD, LCLs

Table 1.1. EBV latencies types and latent gene expressions. The table shows the expression of the genes (indicated by +) in different types of latencies associated with EBV. Note, all lytic infections of EBV generally start with the expression of all latent genes (latency III programme) as exemplified by infectious mononucleosis, Post transplant lympho proliferative disorder (PTLD) and lymphoblast cell lines (LCLs).

fails to completely purge the virus from the host as EBV hides in resting memory B cells which then act as a persistent reservoir for the virus upon reactivation from the latent state. To date the diseases associated or proposed to be associated with EBV could be broadly classified into two categories: non malignant diseases and malignant diseases.

1.4.1. EBV associated non malignant diseases

Most common non malignant diseases known or proposed to be associated with EBV are: infectious mononucleosis, chronic active infection, oral hairy leukoplakia and multiple sclerosis.

EBV infection in the post adolescent can result in infectious mononucleosis (approximately one third of the infections), a self limiting lymphoproliferative disease marked by the latency III programme and lytic infection. Clinically, the patients recover from the disease without any recurrence or developing any severe pathology, however, complications like splenic infarction, airway obstruction and neurological problems have been observed (reviewed in Odumade *et al.*, 2011). Chronic active infection of EBV is characterized by chronic or recurrence of infectious mononucleosis like symptoms. The clinical hallmarks of the disease are abnormally high titre of the EBV antibodies, splenomegaly and/or persistent hepatitis, interstitial pneumonia and lymphadenitis (Kimura *et al.*, 2001). Oral hairy leukoplakia is another form of persistent primary infection of EBV that almost exclusively affects HIV infected individuals (Reichart *et al.*, 1989). The disease is characterized by the extensive replication of EBV particles in oral epithelial cells (Green span *et al*., 1985) with hyperkeratotic and squamous epithelial cell lesions present on the lateral side of tongue. The association of EBV with multiple sclerosis, an autoimmune disease characterized by the depletion of the myelin sheath of neurons, is not currently understood. However, some lines of the evidence point to EBV being a causal and/or contributing factor in the pathogenesis of multiple sclerosis. These include increased incidence of multiple sclerosis among individuals with prior EBV infection, and elevated levels of EBNA1 antibodies before the onset of multiple sclerosis (DeLorenze *et al*., 2006; Levin *et al*., 2010). While EBV does not infect neurons or Schwann cells it is thought that EBV has an effect upon the immune system that promotes this disease. As such, it may also exacerbate multiple other autoimmune disorders.

1.4.2. EBV associated malignant diseases

EBV was the first virus to be associated with human cancer, specifically Burkitt's lymphoma (Epstein *et al*., 1964). Since then several human malignancies have been linked with EBV. The most highly EBV-associated malignancies are: endemic Burkitt's lymphoma (BL), Hodgkin's lymphoma (HL), post transplant lymphoproliferative disease (PTLD), nasopharyngeal carcinoma (NPC) and gastric carcinoma (GC).

BL is a very aggressive B cell lymphoma with three clinical variants: endemic BL, sporadic BL and immunodeficiency associated BL. Each variant differs in their association with EBV, endemic BL has 100% positivity with EBV, while sporadic BL and immunodeficiency associated BL range between 5%-80% and 25%-40% association with EBV respectively (Blum *et al*., 2004). All forms of BL involve translocation of the oncogene, c-myc, to the regulatory region of immunoglobulin gene elements, thereby deregulating c-myc expression (Gutierrez *et al*., 1992).

HL is another commonly associated B cell lymphoma with EBV. Clinically, HL is differentiated into two main variants: non nodular lymphocyte predominant HL (NHL) and classical HL (CHL). NHL accounts for only 5% of the total cases of HL and is usually not associated with EBV. CHL is variably associated with the EBV (10%-90%) depending on the geographical location and co existence of other diseases (Gandhi *et al*., 2004; Jarrett *et al*., 2005).

Post Transplant Lymphoproliferative Disease (PTLD) affects cumulatively 3% of the patients that have undergone transplantation (1% hematopoietic and 2% solid organ transplantation). With around 80% of PTLD being EBV associated, the disease may take 5 months (hematopoietic transplantation) to 5 years (solid organ transplantation) to develop (reviewed in Maeda *et al*., 2009).

NPC is an epithelial carcinoma and on the basis of keratinisation and differentiation of the affected region it is further categorized into three types: differentiated non keratinizing NPC, undifferentiated non keratinizing NPC and keratinizing squamous cell carcinoma. Both types of non keratinizing NPC are totally associated with EBV (100%). As abnormally high titre of IgG and IgA antibodies to the viral capsid antigen (VCA) and early antigen (EA) can be used to predict the onset of NPC (Henle and Henle, 1976), it has been proposed that EBV reactivation and subsequent replication may contribute to the development of NPC. Additionally, since EBV DNA in the NPC tumours appears to be clonal (based on the terminal repeats), it suggests the contribution of proliferating latently infected cells in the development of NPC (Raab-Traub and Flynn, 1986). A caveat to this conclusion is that Moody *et al*. (2003) demonstrated that cells with EBV episomes containing fewer terminal repeats proliferate much faster than cells with longer terminal repeats, suggesting that the previously observed clonality of cells may be the result of selection, rather than evidence of EBV presence at the origin of the tumour.

EBV associated GC is the highest in terms of worldwide incidence amongst all the EBV associated cancers, however, EBV association is lower in this cancer (between 5.2% to 16.0% (van Beek *et al*., 2004)). Clinically, EBV associated gastric carcinoma is now considered as a distinct molecular and pathological entity (Fukayama *et al*., 2001; Ojima *et al*., 1996).

1.5. EBV latent genes

The six EBNA proteins can be expressed from differentially spliced mRNAs, initiated at one of the two promoters (Cp or Wp) that span more than half of the viral genome. Additionally, EBNA1 can be expressed from message initiated at Fp or Qp promoters (Figure 1.2). LMP2 is expressed in the same direction as the EBNAs, but the mRNA spans the terminal repeats and can therefore only be expressed from the episomal genome. LMP1 is expressed in the opposite orientation, using the same promoter as LMP2B. EBERs are the non coding double stranded structural RNAs present just upstream to *OriP*, whereas other microRNAs are the present upstream to LMP1 and LMP2A (Figure 1.2).

1.5.1. EBNA1

EBNA1 is the only protein coding latent gene of EBV which is expressed in three major programmes of latency (latency I-III). Its main role is to ensure propagation and segregation of the viral genome in latently infected dividing cells. EBNA1 structure and functions will be described in detail in section 1.6.

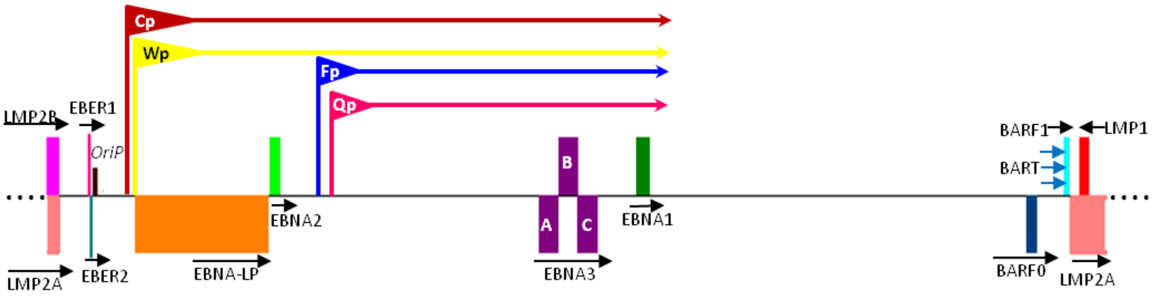


Figure 1.2. Structure of EBV genome. A schematic linear representation of 172 kb circular EBV genome is shown. The latent genes, promoter regions and *OriP* site are shown. Terminal repeats are represented by dots. Promoters are represented by flags with the arrow indicating the transcription direction. For simplicity only three BARTs (micro RNAs clusters; light blue arrows) are shown.

1.5.2. EBNA2

EBNA2 is a transcriptional activator of two viral genes (*LMP1* and *LMP2s*) and cellular genes (CD21, CD23 and others) (Gross *et al*., 2012; Kieff and Rickinson, 2007; Wang *et al*., 1990a; 1990b). EBNA2 interacts indirectly with DNA at the EBNA2 responsive elements (ER2Es) located within the promoter of the LMPs and several cellular genes via DNA binding proteins including Jk-recombination binding protein (RBP-Jk) (Grossman *et al*., 1994). Only recently it has been demonstrated that nuclear chaperone nucleophosmin (NPM1) also plays a critical role in escorting EBNA2 to the promoter region of LMPs (Liu *et al*., 2012). EBNA2 regulates the expression of several viral and host genes including LMP1 and *c-myc* genes pointing to the role of EBNA2 in B cell survival, which in turn facilitate prevalence of the virus in the infected host (Kaiser *et al*., 1999).

1.5.3. EBNA3 gene family

The EBNA3 gene family comprise three related protein: EBNA3A, EBNA3B and EBNA3C *orfs* located tandemly in the EBV genome (Sample *et al*., 1990). The encoded proteins contain seven repeats of leucine, isoleucine or valine that may enable those proteins to dimerise. Their expression results when viral transcription switches to the C promoter (Cp) from the W promoter (Wp) (Gahn and Sugden, 1995). EBNA3A and EBNA3C have been shown to be essential for the B cell transformation (Tomkinson *et al*., 1992; 1993) and growth maintenance of LCLs (Maruo *et al*., 2003; 2005). All EBNA3 proteins competitively inhibit the binding of EBNA2 with RBP-Jks and consequently DNA association (Waltzer *et al*., 1996; Zhao *et al*., 1996) thereby negatively regulating the expression of several viral and host genes which are positively regulated by EBNA2. Conversely, EBNA3C can act as transactivator as its expression in Raji cells (which have the coding exon of EBNA3C deleted) increases the expression of LMP1 (Allday and

Farrell, 1994). Recently White *et al*. (2010) using microarrays reported that over 1000 genes regulated by EBV, require one of the three EBNA3s.

1.5.4. EBNA-LP

Encoded by the leader mRNA sequence of EBNAs in a bicistronic message, EBNA-LP is a protein with repeats of 66 amino acids (the W_1W_2 domains) and a unique 45 amino acid C-terminal sequence (the Y_1Y_2 domains) (Kief and Rickinson, 2007; Sample *et al*., 1986). Along with EBNA2, EBNA-LP induces the G_0 to G_1 transition of resting B cells (Sinclair *et al*., 1994). EBNA-LP is also involved in co-up regulating the target genes of EBNA2, including LMP1 (Nitsche *et al*., 1997). Interactions of EBNA-LP with p53 and retinoblastoma protein (pRb) have also been demonstrated (Szekely *et al*., 1993). Although EBNA-LP is not essential for the immortalisation of the B cells, mutant EBV deleted for EBNA-LP shows an impaired ability to transform B cells. Moreover, deletion of W repeats (to 5 or less) also reduces the transforming ability of the virus (Tierney *et al*., 2011).

1.5.5. LMP1

LMP1 is considered to be the main transforming protein of EBV as it acts as a potent oncogene in several cell types in culture including B cells (Kaye *et al*., 1993; Wang *et al*., 1985). *In vivo*, using LMP1 transgenic mice, the expression of LMP1 in the epithelial cells leads to the onset of early stage of epithelial hyperplasia and this can progress to neoplasia (Stevenson *et al*., 2005; Wilson *et al*., 1990). Under the control of an IgH enhancer, LMP1 expression leads to the development of B cell neoplasia (Hannigan *et al*., 2011). Structurally, LMP1 resembles a tumour necrosis factor family member, CD40. The protein contains three main domains: 1) a short N-terminal cytoplasmic domain, 2) six transmembrane helices and 3) a long C-terminal domain which incorporates three functionally active regions termed C-terminal activating regions (CTAR) 1, 2 and 3 (Eliopoulos and Young, 2001; Li and Chang, 2003). CTAR1 and CTAR2 interact with tumour necrosis factor receptor associated factors (TRAFs) and tumour necrosis factor-receptor death domain proteins (TRADDs) respectively. Such interactions instigate several signalling pathways including the NFκB pathway (Huen *et al*., 1995), MAPK pathways (Eliopoulos *et al*., 1999) and P13K/Akt pathways (Dawson *et al*., 2003). Through these pathways, LMP1 deregulates the expression of multiple genes including EGFR (Kung *et al*., 2011) and EGFR ligand (Hannigan et al., 2007), apoptotic protein Bcl-2 (Henderson *et al*., 1991) and stimulates inflammatory cytokine production in LMP1 transgenic mice (Hannigan *et al*., 2010; 2011).

1.5.6. LMP2A/2B

LMP2A and 2B are encoded by a single gene *LMP2*, which is intervened by the terminal repeat sequences, the point at which linear ends of EBV genome join to form an episome (Sample *et al*., 1989). Both genes are expressed simultaneously under the control of two promoters located 3kb apart. LMP2A and LMP2B are identical except for the presence of an additional 5' exon in LMP2A giving it an additional 119 amino acids, N-terminal domain compared to LMP2B (Longnecker and Kieff, 1990). However, the 12 transmembrane domains are common to both LMP2A and 2B. Analysis of EBV recombinant mutants deleted for LMP2 showed that both these proteins are not essential for B cell transformation (Longnecker *et al*., 1993a; 1993b; 2000). Two out of the eight tyrosine residues present at the N-terminal domain constitute a immunoreceptor, tyrosine-based activation motif (ITAM), that play a central role in the proliferation and differentiation of lymphocytes by interacting with the protein kinases of Src and Syk families. It has been proposed that LMP2A can block normal B cell receptor (BCR) signalling by assembling the tyrosine kinase to its ITAM (Fruehling and Longnecker, 1997). This finding is further supported by studies conducted using LMP2A transgenic mice. LMP2A expressing B cells from these transgenic mice survive without producing immunoglobulins suggesting that LMP2A may facilitate survival of B cells, even in the absence of essential BCR signalling suggesting that LMP2A can provide the survival signal that would otherwise be transduced by the BCR (Caldwell *et al*., 1998).

1.5.7. EBERs

The EBERs (EBER1 and EBER2) are non polyadenylated, non coding polIII RNAs, which along with EBNA1 are consistently expressed in all programmes of EBV latencies (with some exceptions like in GC). EBERs are amongst the most highly expressed viral genes in latency, therefore routinely used in the diagnosis of EBV infected cells (Chang *et al*., 1992). Sequentially, EBERs are highly conserved among different EBV isolates and are thought to adopt a secondary structure conformation. The secondary structure contains stem loops which are suggested to interact with several proteins including protein kinase R (PKR) (Takada and Nanbo, 2001). Binding of EBERs with PKR is thought to inhibit interferon responses and in turn apoptosis, providing EBV with an arsenal to counter the host innate immune response (Nanbo *et al*., 2002). EBER expression in EBV negative Akata cell line results in partial restoration of tumourogenic phenotype of EBV+ Akata cells (Ruf *et al*., 2000). EBERs also confer resistance to apoptosis when expressed in intestinal epithelium cells, via blocking PKR activity (Nanbo *et al*., 2005). Moreover, EBER deleted mutant of EBV shows 100 fold reduced transforming ability as compared to wild type (Yajima *et al*., 2005). Consistent with these observations, EBER1 expression in transgenic mice led to lymphoid hyperplasia ultimately followed by B cell malignancy (Repellin *et al*., 2010). Another study, based on recombinant viruses (either with EBER1 or EBER2) suggests that the transforming ability of EBERs is due to EBER2 and not EBER1 (Wu *et al*., 2007).

1.5.8. *Bam*H1A region transcripts

This region of the EBV genome encodes highly expressed RNAs termed *Bam*H1A rightward transcripts or BARTs (Karran *et al*., 1992; Smith *et al*., 2000). During latent infection of EBV, 29 miRNAs are expressed from three clusters in the EBV genome, of these two clusters are made from BARTs (Edwards *et al*., 2008). Other protein coding transcripts from the same region include BARF0 and BARF1. BARF1 encodes a 31 kDa protein which was originally considered to be a lytic gene but also has been found to be expressed in NPC and GC (Decaussin *et al*., 2000, zur Hausen *et al*., 2000). Furthermore, BARF1 is a potential oncogene as its expression leads to the transformation of rodent fibroblasts in culture and it can induce expression of the antiapoptotic gene Bcl2 (Sheng *et al*., 2001).

1.6. EBNA1 in depth

EBNA1 is the only protein coding latent gene that is expressed in all latencies of EBV. EBNA1 was first described in fresh tumour biopsies of NPC (de The *et al*., 1973) and subsequently identified in many EBV infected tissues (Wright *et al*., 1975; Yamamoto *et al*., 1975). EBNA1 is an 88 kDa, 641 amino acid containing protein. More than one third of the protein is composed of a glycine and alanine repeats region, which were first identified using antibodies present in the human sera in 1983 by Hennessey and Kieff. EBNA1 is a pleiotropic protein and is involved in a variety of functions, including genome maintenance, transactivation, resistance to apoptosis and oncogenesis. The diversity in EBNA1 functions is primarily due to its ability to interact with several host and viral biomolecules (Figure 1.3; 1.4).

1.6.1. The role of EBNA1 in genome replication and maintenance

The main biological role of EBNA1 in the virus is to facilitate the non random segregation of viral genomes in latently infected cells. During latent EBV infection, the circular episome of the virus undergoes one round of bi directional replication per cell division. EBV replication in latent infection starts at unique region, *oriP* and while using only one viral protein, EBNA1, it relies heavily on the host replication machinery. *oriP* is composed

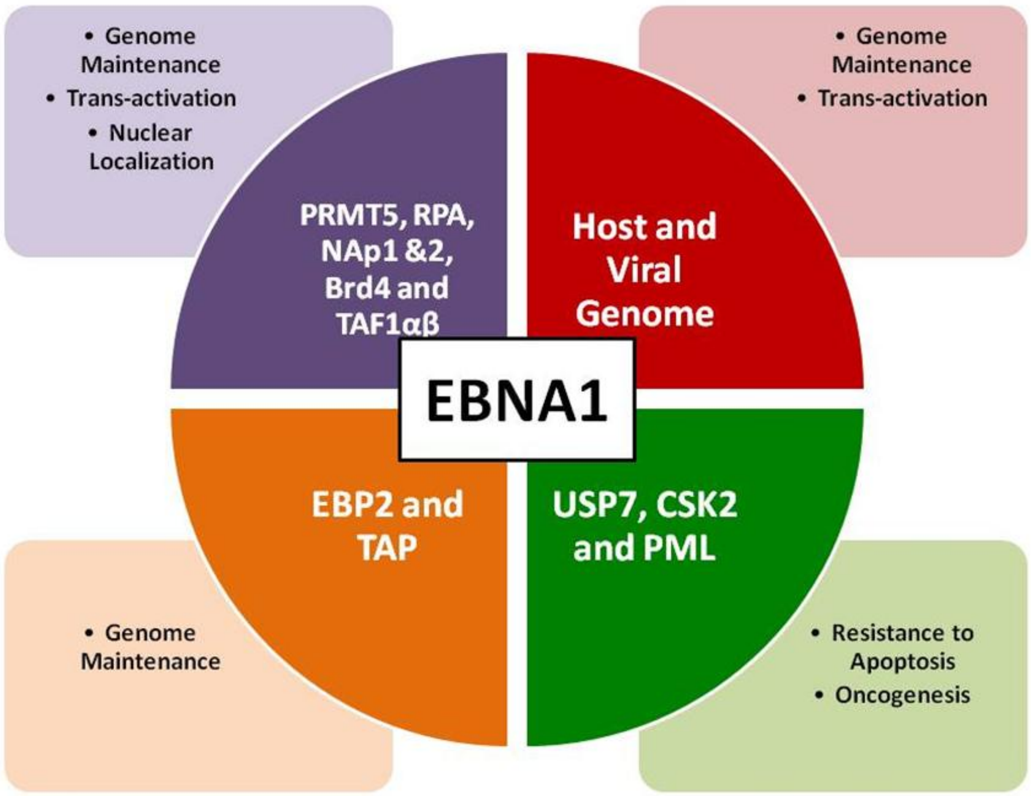


Figure 1.3. Binding partners and functions of EBNA1. The figure represents the pleiotropic nature of EBNA1 involved in a variety of functions including genome maintenance, transactivation and resistance to apoptosis owing to its binding with multiple host proteins as well as DNA and RNA.

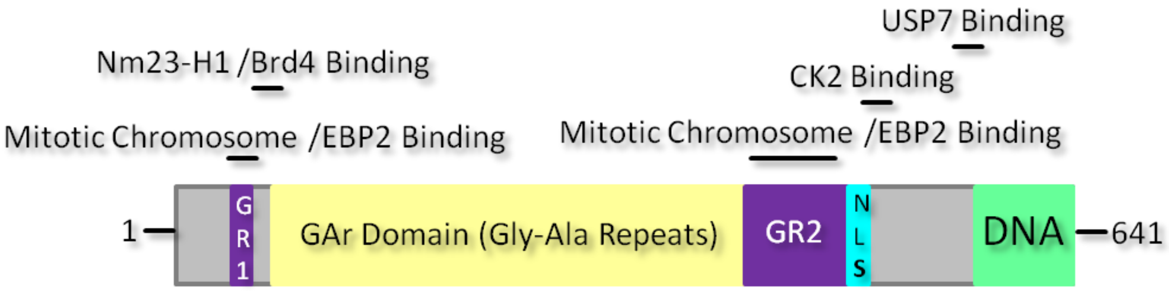

USP7 Binding
Nm23-H1 /Brd4 Binding
CK2 Binding
Mitotic Chromosome /EBP2 Binding
Mitotic Chromosome /EBP2 Binding
1
GR1
GAr Domain (Gly-Ala Repeats)
GR2
NLS
DNA
641

Figure 1.4. Protein interacting regions and domain distribution of EBNA1. Schematic representation of the EBNA1 protein is shown. Different structural/functional domains are coloured differently and indicated. The horizontal bars represent the position and span of different protein binding regions on EBNA1 sequence. Note, EBNA1 interacts with multiple host proteins, binding region of only few are mapped and shown here.

of two functional elements: a Dyad symmetry (DS) element and a family of repeats (FR), situated 1kbp apart. The DS element contains 65bp central dyad symmetrical sequences flanked by 3 copies of a 9bp sequence (nonamers) at each end. The central dyad symmetrical region contains 2 of the 4 binding sites for EBNA1 and optimal replication of EBV episomes requires all four sites of EBNA1 interaction and the nonamers. Moreover, the space (3bp) between two adjacent EBNA1 binding sites is also crucial for effective replication of the viral genome. The FR repeats are 20 tandem copies of a 30 bp sequence, each of which contains an 18bp region for EBNA1 binding (Figure 1.5). Interaction of EBNA1 with both DS and FR is critical for EBV genome replication (Reismann *et al*., 1985; reviewed in Frappier, 2012). Paradoxically, binding of EBNA1 with FR may also inhibit the replication by resisting the unwinding of DNA and movement of the replication fork, limiting replication to once/S phase (Dhar and Schildkraut, 1991).

The co-crystallized structure of an EBNA1 C-terminal dimer with DNA has shown the presence of two important regions, a DNA recognition helix (461-503 a.a) that flanks the second dimerisation or core domain (504-604 a.a). Despite the limited sequence homology, the core domain of EBNA1 shares noticeable structural similarity with the dimerisation domain of the E2 protein of papilloma virus (Bochkarev *et al*., 1995; 1996). The EBNA1 homodimer recognizes and binds to the 18bp pallindromic sequence present in DS and FR (Ambinder *et al*., 1990; Rawlins *et al*., 1985). This DNA-protein interaction is mediated by the C-terminal domain (459-603 a.a) of EBNA1. However, it is now increasingly evident that in addition to the C-terminal one third of the protein, N- terminal region of EBNA1 also contribute to viral DNA replication and genome maintenance (Deng *et al*., 2005; Holowaty *et al*., 2003; Shire *et al*., 1999). Binding of EBNA1 with USP7 has been shown to increase its efficiency to bind with DNA *in vitro*. Depletion of cellular USP7 negatively affects the EBNA1 binding to *oriP* (Sarkari *et al*., 2009; 2010). Conversely, higher DNA replication activity has been shown in the presence of an EBNA1 mutant lacking the USP7 binding region, suggesting that EBNA1-USP7 binding may negatively regulate viral DNA replication (Holowaty *et al*., 2003).

EBV genome replication during latent infection also involves extensive recruitment of the host replication machinery (reviewed in Frappier, 2012). Studies have shown that the host cell origin recognition complex (ORC) and minichromosome maintenance (MCM) complex are recruited on the DS element at *oriP*, suggesting their role in EBV genome replication (Chaudhuri *et al*., 2001; Dhar *et al*., 2001). Consistent with this it has been demonstrated that cells having a mutation in ORC fail to stably replicate the EBV genome (Dhar *et al*., 2001). Recruitment of ORC to the DS region is mediated by EBNA1 via its

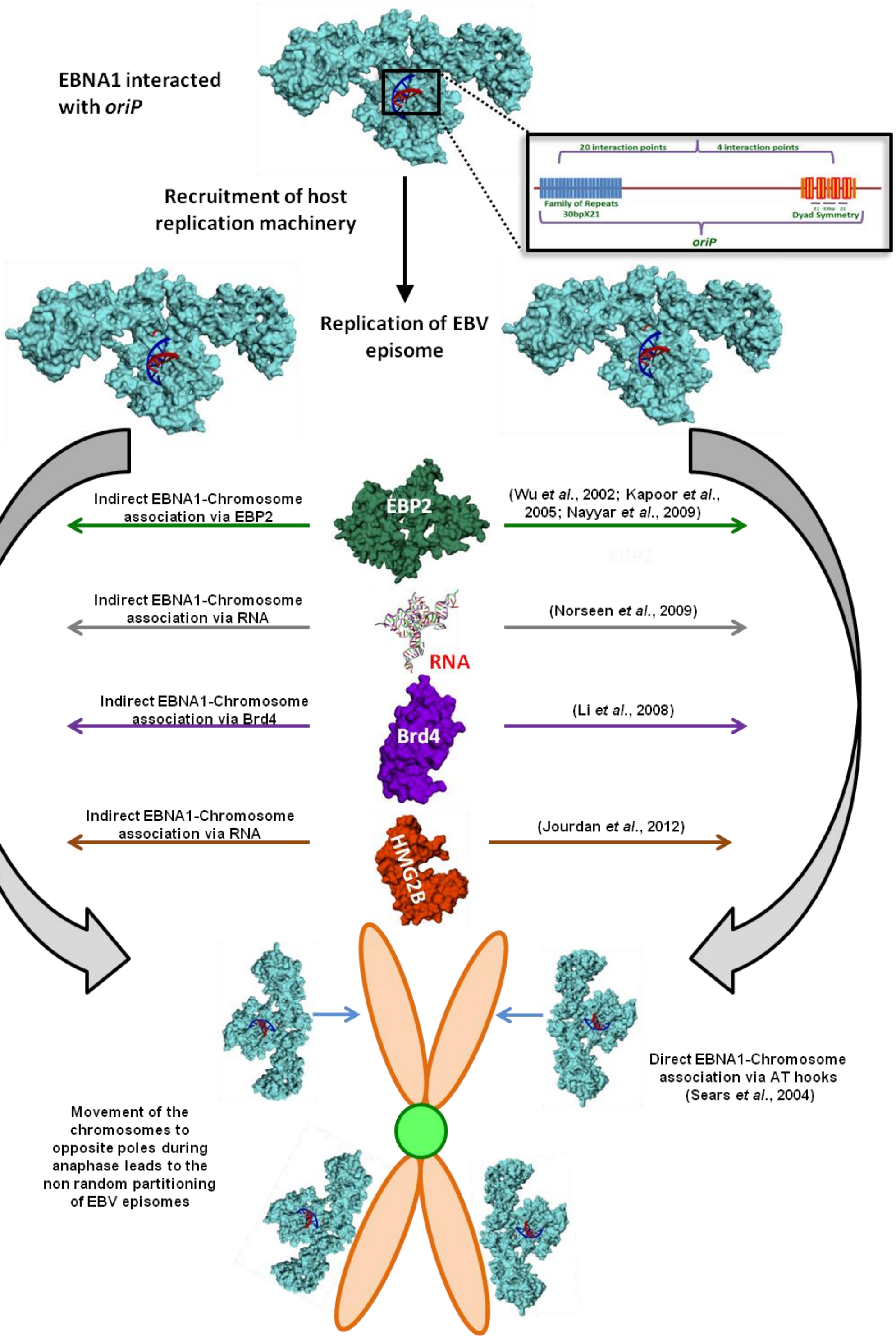

EBNA1 interacted with *oriP*
20 interaction points
4 interaction points
Family of Repeats
30bpX21
21 33bp 21
Dyad Symmetry
oriP
Recruitment of host replication machinery
Replication of EBV episome
Indirect EBNA1-Chromosome association via EBP2
EBP2
(Wu *et al*., 2002; Kapoor *et al*., 2005; Nayyar *et al*., 2009)
Indirect EBNA1-Chromosome association via RNA
RNA
(Norseen *et al*., 2009)
Indirect EBNA1-Chromosome association via Brd4
Brd4
(Li *et al*., 2008)
Indirect EBNA1-Chromosome association via RNA
HMG2B
(Jourdan *et al*., 2012)
Direct EBNA1-Chromosome association via AT hooks (Sears *et al*., 2004)
Movement of the chromosomes to opposite poles during anaphase leads to the non random partitioning of EBV episomes

Figure 1.5. EBNA1 role in genome maintenance. Schematic representation of functional components of EBV genome maintenance is shown. To date, a scientific consensus is lacking regarding the molecular mechanism underlying non random partitioning of the EBV episome in dividing latently infected cells. Shown here are all the proposed mechanisms with the corresponding references. Inset: the structure of *oriP* is shown with both DS and FR repeat elements, the main DNA binding site of EBNA1. Genome maintenance role of EBNA1 is derived from its direct interaction with metaphase chromosome or indirect interaction with chromosome by EBP2, RNA, Brd4 and HMG2B as indicated. Surface topologies of the protein molecules are shown. Predicted models of EBP2 and EBNA1 (this study) are used to represent the surface of full length molecules. Partial structures of Brd4 (PDB id: 4HXP) and HMG2B (PDB id: 1J3C) are used to represent the respective proteins.

interaction with RNA and/or Cdc6 protein (Moriyama *et al*., 2012; Norseen *et al*., 2008). As marking of the origin of replication is mediated by EBNA1, it has been proposed that assembly of ORC at the same position may serve some additional and/or different role in EBV genome replication (Frappier, 2012). The Minichromosome maintenance (MCM) complex is also recruited at the *oriP* via two accessory proteins, Cdc6 and Cdt1 and may undertake a helicase action in EBV episome replication (Dhar *et al*., 2001). Another cellular protein recruited at the *oriP* is telomere repeat binding factor 2 (TRF2), which interacts with the three nonamer repeats at DS mostly during the G1/S phase (Deng *et al*., 2002). Although the N-terminal domain of TRF2 was shown to be supportive in order to recruit ORC to DS, its absolute requirement is still unclear (Atanasiu *et al*., 2006) as a later study by Moriyama *et al*., 2012 showed that EBNA1 mediated recruitment of ORC and Cdc6 at *oriP* is TRF2 independent. Another study has demonstrated that depletion or deletion of TRF2 advances the EBV episomal replication from late S to mid S phase, possibly by recruiting histone deacetylase (HDC) 1 and 2 (Zhou *et al*., 2009). TRF2 has also been shown to recruit ChK2 to DS during the G1/S phase (Zhou *et al*., 2010) and recombination proteins like MRE11 and NBS1 during S phase (Dheekollu *et al*., 2007). Telomere associated factors such as TRF1 also bind to the DS of *oriP*, however its role in viral genome replication is unclear. Two other proteins have also been found to be recruited to the DS: tankyrase poly (ADP-ribose) polymerase (TPP) and hARP (Deng *et al*., 2002; 2003b). TPP also directly binds to EBNA1 and this interaction negatively regulates DNA replication, as mutation in the TPP binding site of EBNA1 (Gly81 and Gly425) leads to an increase in *oriP* mediated DNA replication (Deng *et al*., 2005). By contrast depletion of hARP results in a decrease in *oriP* mediated EBV episome replication (Deng *et al*., 2003). Timeless (tim) and tipin are two other proteins which are recruited to *oriP*. Decrease *oriP* mediated replication and increased double strands breaks have been observed at *oriP* when tim protein is depleted from the cell. Additionally, both timeless and tipin are known for stabilizing replication forks at repetitive sequences (Dheekollu *et al*., 2011). EBNA1 also directly interacts with template activating factor Iβ (TAFIβ) via Gly-Arg region (325-376 a.a) of EBNA1 and recruits it to *oriP*. Recruitment of TAF1β negatively regulates viral DNA replication by assembling histone acetylase and HDA, which in turn modify the chromatin structure (Wang and Frappier, 2009).

Although EBNA1 association with the viral episome has been explored in detail, the mechanism of EBNA1 association with host chromosomes, for effective segregation of the EBV genome, is still under debate (Figure 1.5). Nevertheless, based on observations taken from florescent microscopy, deletion mutants, immuno precipitation and biochemical fractionation, it seems that EBNA1 interacts with human chromosomes at AT rich regions via the two EBNA1 Gly-Arg regions LR1 (33-89) and LR2 (328-378), therefore termed AT hooks. This interaction may be direct (Sears *et al*., 2004) or indirect (in later stages of cell cycle) by a nucleolus protein p40 or EBP2 (Kapoor *et al*., 2005; Nayyar *et al*., 2009; Shire *et al*., 1999; Wu *et al*., 2002). Replacement of LR1 and/or LR2 with a similar region from the high mobility group AT hook 1 protein, was as effective in the segregation of the EBV episome, supporting the idea of direct association of EBNA1 with human chromatin (Sears *et al*., 2004). By contrast, depletion of EBP2 (by silencing or dissociation by aurora kinase) notably decreases the association of EBNA1 with host chromosomes (Kapoor *et al*., 2005), supporting the idea that EBNA1 interacts with metaphase chromosome indirectly through EBP2. Similarly, colocalization of EBNA1 and EBP2 was observed on chromosomes from metaphase to telophase (Nayyar *et al*., 2009). Moreover, Jourdan *et al.* (2012) has demonstrated that the interaction of EBP2 and EBNA1 occurs during interphase and not in the later stages of mitosis. These authors purposed a non obligate loading role of EBP2, for EBNA1 association with metaphase chromosomes. In the same study an alternate binding partner, HMGB2, was proposed, for stabilizing the EBNA1-chromosome interaction. Adding further complexity is the observation that Braco-19, a G-quardi duplex RNA disruptor, also inhibits EBNA1 association with cellular chromosome (Norseen *et al*., 2009), suggesting a role for RNA in EBNA1-chromosome association. Additionally, it has recently been shown that the relocalization of EBNA1 during the cell cycle (dispersed throughout nucleus to metaphase chromosomes) also depends on Gly-Ala repeat (GAr) length (Coppotelli *et al*., 2013). Finally, by using a reconstituted virus replication system in yeast and EBNA1 deletion mutation analysis, the direct physical interaction between EBNA1 and Brd4 was shown and an additional but dispensable mechanism for EBV episome tethering to the metaphase chromosome was proposed (Lin *et al*. 2008). Nevertheless, association of EBNA1 with metaphase chromosomes subsequently leads to the non random segregation of EBV episomes into the daughter cells, as the chromosomes move towards the opposite poles during anaphase (Figure 1.5).

1.6.2. EBNA1 as transactivator

EBNA1 acts as a transcriptional transactivator for both viral and cellular genes. Upon binding to FR in the EBV genome, EBNA1 mediates the transcription of several latent genes including LMP1 (Gahn and Sugden, 1995). Paradoxically, EBNA1 specific binding to viral promoter Qp has shown to negatively regulate its own transcription (Sample *et al*., 1992; Sung *et al*., 1994). Although the molecular mechanism of EBNA1 transactivation activity is not clear, some details are available. For example EBNA1 binding to *oriP* may result in the loop formation between FR and DS and this structure has direct consequences upon viral replication and latent gene transcription (Frappier and O'Donnell, 1991; Su *et al*., 1991). It may induce re-organization of the chromatin structure which results in the progressive recruitment of additional transcriptional machinery to the region (Niller and Minarovitis, 2012). Moreover, DNA looping may further be induced by binding partners of EBNA1, such as Brd4, PRMT5, TAF1 and NAF1 (Lin *et al*., 2008; Malik-Soni and Frappier, 2012; Wang and Frappier, 2009). Additionally, a region between 64-89 a.a. in EBNA1 contains two strongly conserved cysteine residues which might bind with Zn. Substitutions at these cysteine residues or chelation of cellular zinc impairs the transactivation ability of EBNA1 (Aras *et al*., 2009).

In addition to the viral episome, it is now widely established that EBNA1 specifically interacts with the host cellular DNA. This interaction may have consequences on the expression of cellular genes. Using chromatin immunoprecipitation (ChIP) supported by promoter array and/or deep sequencing analyses, several cellular sites have been identified as targets for EBNA1 binding. Interestingly, these sites are diverse in terms of their sequence, thus unlike the viral sequence, a clear consensus is still lacking regarding the target sequence(s) in the human genome for EBNA1 binding (Canaan *et al*., 2009; Lu *et al*., 2010). Moreover, the functional significance of this interaction is elusive. For example, EBNA1 high affinity binding to FR like elements in the human genome has been demonstrated but without any evidence of ORC or MCM protein recruitment (d'Herouel *et al*., 2010; Lu *et al*., 2010). It is possible that lack of such recruitment is due to the requirement of a 21 bp span between two adjacent EBNA binding site for recruiting cellular proteins involved in replication (Bashaw and Yates, 2001). Similarly, binding with the FR like elements present on human chromosome 11 does not show any alteration in the transcriptional activity of the nearby genes (Lu *et al*., 2010). Consistent with these studies, cellular promoters to which EBNA1 shows an affinity, when cloned upstream of the luciferase gene, have not shown any alteration in transcription in the presence of EBNA1 (Dresang *et al*., 2009). Conversely, exogenous expression of EBNA1 enhanced the transcription of several genes including survivin in the EBV negative BL cell lines DG75 and BJAB. Decreased transcription has also been observed in a similar set of genes when EBNA1 is depleted in the EBV positive Raji cell line using siRNA (Lu *et al*., 2010; Canaan *et al*., 2009). Moreover, the presence of EBNA1 results in the two and four fold increase in the expression of *ATF2* and *c-Jun* genes respectively in NPC cells (O'Neil *et al*., 2008).

1.6.3. EBNA1 as oncogene

Given that EBNA1 is the only latent protein expressed in BL and is consistently expressed in all latency programmes of EBV, it has been speculated that EBNA1 may contribute to EBV associated oncogenesis. Considerable amount of evidences has been accumulated pointing to a direct role of EBNA1 in EBV associated oncogenesis. The first demonstration that EBNA1 might have oncogenic activity was made using transgenic mice (Wilson *et al*., 1996; Wilson and Levine, 1992). These studies showed that two independent, EBNA1 expressing transgenic mouse lines were predisposed to B-cell lymphoma. Although this finding was against the dogma at the time (and another group failed to show EBNA1 mediated transgenic oncogenesis (Kang *et al*., 2001; 2005; 2008), numerous studies since then have pointed to EBNA1 having an oncogenic activity, particularly in increasing the cell survival. Increased immortalization has been observed in EBV infected B lymphocytes due to the presence of EBNA1 (Altmann *et al*., 2006; Humme *et al*., 2003). Additionally, expression of EBNA1 dominant negative mutants decreased cell survival and increased apoptosis in BL cells (Kennedy *et al*., 2003). In line with this, silencing of EBNA1 increased the survival in BL and NPC cell lines (Hong *et al*., 2006; Yin *et al*., 2006). This was also found *in vivo* using EBNA1 expressing transgenic mice. EBNA1 lymphocytes showed a prolonged survival, but this was dependent in culture on the supplement of IL2. In the same study increased expression of bcl-XL and recombination activating genes (RAG 1 and 2) was also noticed (Tsimbouri *et al*., 2002). In another study a synergistic effect of *Myc* and EBNA1 leading to the early onset of the lymphomogenesis was observed in the transgenic system (Drotar *et al*., 2003). Furthermore, increased primary tumour formation and metastasis have been observed in response to EBNA1 expression in the HONE1 NPC cells (Sheu *et al*., 1996), breast cancer cells (Kaul *et al*., 2007) and gastric carcinoma cells (Cheng *et al*., 2010).

At the molecular level, evidence to delineate the underlying mechanism of cell survival and consequently oncogenesis involving EBNA1 is accumulating. Among these are: destabilization of p53 (Holowaty *et al*., 2003; Saridakis *et al*., 2005), destabilization of promyelocytic leukemia (PML) nuclear bodies (Sivachandran *et al*., 2008; 2012), induction of reactive oxygen species (ROS) (Cao *et al*., 2012; Gruhne *et al*., 2009a) and modulation of signalling pathways (Wood *et al*., 2007; Valentine *et al*., 2010) (Figure 1.6).

Direct binding between EBNA1 (436-450 a.a) with the MATH/TRAF domain of USP7 has been demonstrated *in vitro* (Holowaty *et al*., 2003; Saridakis *et al*., 2005). USP7, a deubiquitinase, removes the polyubiquitin chain from the key tumour suppressor protein p53, as well as its ubiquitin E3 ligase MDM2, thereby protecting both proteins from proteosomal degradation and promoting their stabilization (Li *et al*., 2004; Hu *et al*., 2006).

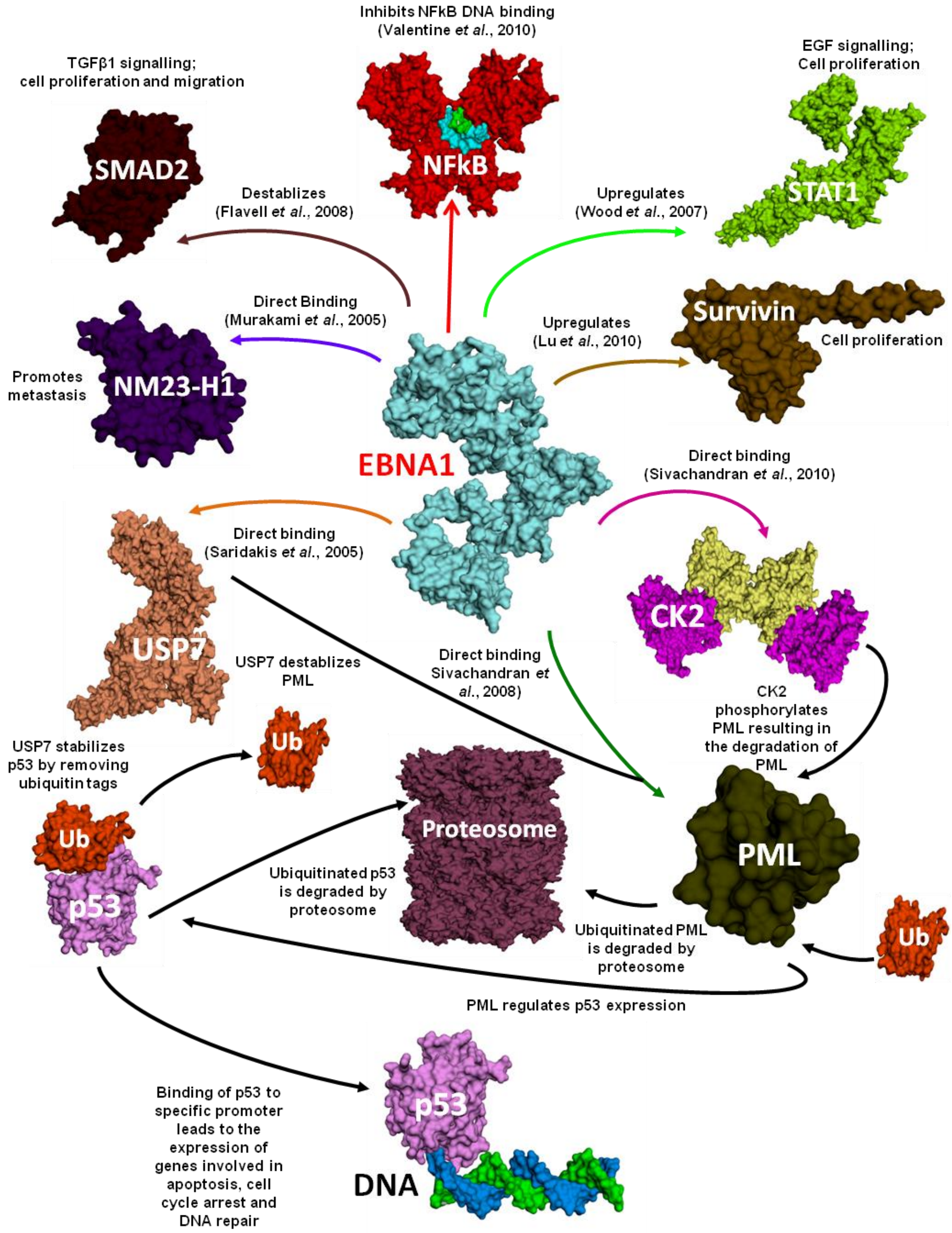

Inhibits NFkB DNA binding (Valentine et al., 2010)
TGFβ1 signalling; cell proliferation and migration
EGF signalling; Cell proliferation
SMAD2
NFkB
STAT1
Destablizes (Flavell et al., 2008)
Upregulates (Wood et al., 2007)
Direct Binding (Murakami et al., 2005)
Upregulates (Lu et al., 2010)
Survivin
Cell proliferation
Promotes metastasis
NM23-H1
EBNA1
Direct binding (Sivachandran et al., 2010)
Direct binding (Saridakis et al., 2005)
USP7
USP7 destablizes PML
Direct binding Sivachandran et al., 2008)
CK2
CK2 phosphorylates PML resulting in the degradation of PML
USP7 stabilizes p53 by removing ubiquitin tags
Ub
Proteosome
PML
Ub
p53
Ubiquitinated p53 is degraded by proteosome
Ubiquitinated PML is degraded by proteosome
Ub
PML regulates p53 expression
Binding of p53 to specific promoter leads to the expression of genes involved in apoptosis, cell cycle arrest and DNA repair
p53
DNA

Figure 1.5. EBNA1 as an oncogene: Depiction of models, based on the studies suggesting the role of EBNA1 in cell survival and resistance to apoptosis. Surface topology of all protein and DNA molecules are shown and labelled. The predicted models of full length EBNA1 (cyan) and USP7 (brown) are used. Full length structures of CK2 (purple and yellow PDBid: 1JWH), proteosome (dark purple; PDBid:1G65), ubiquitin (orange; PDBid:1AAR), survivin (brown; PDBid:4AON), NFkB (red; PDBid:1NFK), SMAD2 (dark brown; PDBid: 1KHX), STAT1 (light green; PDBid: 1YVL), p53 (light purple; PDBid:1TUP) and NM23-H1(blue; PDBid: 3L7U) have been retrieved from RCSB protein data bank. PML (green; PDBid: 1BOR) is represented by their partially available structure. Short description between the EBNA1 and the associated proteins and the suggested outcome are indicated on the connecting arrows, coloured as partner molecule.

Biochemical and structural studies have demonstrated that all, EBNA1, MDM2 and p53 compete for the same region for binding with USP7, however EBNA1 shows highest affinity amongst the three. As a consequence it has been proposed that during EBV infection, EBNA1 binds with USP7 and thus interferes with the otherwise tightly regulated ubiquitination/deubiquitination processes of p53; this dys-regulation leads to the proteosomal degradation of p53 (Hu *et al*., 2006; Li *et al*., 2004; Saridakis *et al*., 2005). *In vivo* studies have also demonstrated that expression of EBNA1 in U20S (Saridakis *et al*., 2005) and CNE2 NPC (Sivachandran *et al*., 2008) and GC cells (Sivachandran *et al*., 2012a) results in the depletion of p53. Taken together these observations plausibly suggest that EBNA1 can disturb the steady state levels of p53 in the EBV infected cells to support cell survival. However, treatment of LCL with DNA damaging agents has led to p53 mediated apoptosis (O'Nions *et al*., 2006) and p53 is intrinsically mutated in 50% of the BL cases but not in NPC (Schmitz *et al*., 2012). The role of EBNA1 in the cell survival could in part be explained by its interaction with PML nuclear bodies. PML nuclear bodies consist of nuclear proteins, which are involved in apoptosis and DNA repair (reviewed in Bernardi *et al*., 2007; Salomoni *et al*., 2012). EBNA1 preferentially interacts with one of the five human isoforms of PML, namely PMLIV. Additionally, the EBNA1 region defined (387-394) also directly interacts with the β regulatory subunit of casein kinase 2 (CK2), which phosphorylates PML. CK2 mediated phosphorylation of PML signals for its ubiquitination and consequently proteosomal degradation (Scaglioni *et al*., 2006; 2008; Sivachandran *et al*., 2010). Moreover, EBNA1 expression in both GC and NPC results in the depletion of PML nuclear bodies (Sivachandran *et al*., 2008; 2012). Moreover, the EBNA1-USP7 interaction is thought to be important in this regard, but lacks clarity for the underlying molecular mechanism. However, some studies have demonstrated that USP7 can induce the degradation of PML nuclear bodies, independently of its EBNA1 interaction and its DUBs catalytic activity (Sarkari *et al*., 2011).

Microarray analyses of Ad/Ah cells, with elevated expression of EBNA1, have shown perturbed expression of 162 genes compared to Ad/AH cells infected with a rEBV or in the C666-1 EBV-positive NPC cell line. Among these genes is STAT1, a protein with an established role in apoptosis mediated and independent cell death (Wood *et al*., 2007). In addition, microarray analyses comparing EBNA1 expressing transgenic B cells with controls, also shows elevated STAT expression (Tsimbouri and Wilson, unpublished data). Moreover, a potential EBNA1 binding site is located near the STAT1 transcription initiation site (Dresang *et al*., 2009). Similarly, the presence of EBNA1 in the Ad/Ah cells reduces the half life of SMAD2, an important mediator of TGFβ1 signalling, implicating EBNA1 in the interference of the TGFβ1 signalling cascade. Moreover, expression of βigh3, a gene regulated by TGFβ1 signalling has been found to be reduced in the presence of EBNA1 (Wood *et al.*, 2007). In support of this, another study using HL cells has also shown reduced turnover of SMAD2 and reduced expression of Protein Tyrosine Phosphate Receptor K (PTPRK) in the presence of EBNA1 (Flavell *et al.*, 2008). It is also noteworthy that 15% and 20% of promoters of differentially expressed genes (in the presence of EBNA1) in Ad/Ah cells bear DNA binding motifs for NFκB (Valentine *et al.*, 2010) and AP-1 (O'Neil *et al.*, 2008) respectively. Further studies have revealed that EBNA1 can negatively affect NFκB activity by inhibiting its binding with DNA (Valentine *et al.*, 2010). Taken together, the data show that EBNA1 modulates certain signalling cascades, known for their role in cell survival and apoptosis.

In BL cell lines, expression of EBNA1 increases the production of ROS and consequently genomic instability (Gruhne *et al.*, 2009b). ROS exhibits many cellular effects, not least are genomic instability by DNA damage and induction of apoptosis (Avery, 2011). Similarly, in CNE2 cells, expression of EBNA1 results in the increased expression of Nox2 and production of ROS. Moreover, increased telomeric instability has been observed under the EBNA1 mediated production of ROS (Cao *et al.*, 2012; Gruhne *et al.*, 2009a,b; Kamranvar *et al.*, 2011). This suggests that EBNA1 may cause genomic instability by increasing the production of ROS.

EBNA1 also binds with cellular metastatic inhibitor Nm23-H1 through amino acids 65-89 and the interaction is thought to impair the Nm23-H1 function. In agreement, increased cell migration has been observed in LCL in the presence of EBNA1 (Murakami *et al.*, 2005). Moreover, nucleoproteosomal analysis of NPC cells has shown increased level of Nm23-H1 and two other metastasis associated proteins namely, maspin and stathamin 1 in the presence of EBNA1 (Cao *et al.*, 2012). In addition, increased expression of Nm23-H1 has been correlated with increased expression of apoptotic genes such as caspases 3, caspases 9, Bcl-X and p53 (Choudhuri *et al.*, 2010). Though details are still elusive, but the evidence suggest that by impairing the function of Nm23-H1 through direct binding, EBNA1 may decrease apoptosis in the EBV infected cells. Finally EBNA1 is also known to increase the expression of an anti apoptotic protein, survivin (Lu *et al.*, 2011) which has a role in cell proliferation.

In summary, the direct role of EBNA1 in cell survival and proliferation and conferring resistance to apoptosis and potentially oncogenesis is substantiated by a significant amount of evidence. This suggests that aside from its core function in the virus of genome maintenance, EBNA1 is very likely to be involved in the onset and/or progression of EBV associated cancers.

1.7. Ubiquitin Specific Proteases

To undertake their biological roles, most proteins are required to adopt stable and functionally favourable structural conformations. The process of ubiquitination and deubiquitination plays a pivotal role in ensuring this structural stability, as well as enabling rapid removal of proteins that are no longer required within the cell (reviewed in Amerik and Hochstrasser, 2004; Komander *et al*., 2009; Peng *et al*., 2003). Ubiquitin tagging of a protein is mainly mediated by the sequential activities of three different ligases: E1, E2 and E3. Briefly, three main steps are involved in the molecular cascade of ubiquitination: first an energy mediated process leads to the linkage of two activated ubiquitin molecules to E1 ligase, via thiol ester and adenylate linkages. Second, this thiol linked molecule is transferred to E2 ligase and finally the ubiquitin molecule is transferred from E2 to target protein by interacting with a substrate specific E3 ligase (reviewed in Pickart, 2001). As substrate specificity is mostly defined by the E3 ligases, it is not surprising that there are more variants of E3 ligases in comparison to E1 and E2 ligases. To date, over 650 ubiquitinated proteins and over 600 E3 ligases have been identified using mass spectrometry and genome wide analysis respectively (Li *et al*., 2008; Meierhofer *et al*., 2008). The diversity of E3 ligases is further augmented by the length of the ubiquitin chain attached to the substrate via isopeptide bonds between the carboxy terminus of Gly of the substrate protein and one of seven internal Lys residues (Lys6, Lys11, Lys27, Lys33, Lys 48 and Lys63) of ubiquitin. However, considering the protein diversity in human proteome (from over 20,000 genes), many novel E3 ligases and target proteins may yet to be identified.

Ubiquitination instigates the proteosomal degradation of redundant or improperly folded protein molecules. To counter balance this, properly folded ubiquitinated proteins get untagged by the activity of another set of proteins collectively referred as deubiquitinases (DUBs). Around 100 DUBs have been identified so far which are catalytically active (Komander, 2010; Nijman *et al*., 2005b). DUBs are characterised into 5 different families, namely: Ubiquitin C-terminal hydrolases (UCHs), ubiquitin specific proteases (USPs), ovarian tumour proteases (OTUs), Josephins and JAB1/MPN/MOV34 metalloenzymes (JAMMs) (reviewed in Kommander *et al*., 2009). Of these five families, USPs stands distinct in terms of structural and functional diversity.

To date over 50 USP paralogues in the human genome have been identified. The proteins encoded by these genes vary considerably in size and domain architecture (reviewed in Kommander *et al*., 2009). Akin to the other DUBs, the main function of USPs is to regulate their target protein turnover and this is mediated through the cysteine peptidase activity of the C19 peptidase domain. Structurally, the peptidase domain of USPs consist of three subdomains referred to as palm, thumb and fingers of a hand, with catalytic sites positioned at the interface of all three subdomains. The main interaction between DUBs, (including USPs) and the ubiquitin molecule is established between the ubiquitin binding domain (UBD) of DUBs and Ile44 of ubiquitin (Zhu *et al*., 2007). Additionally, some USP domains are structurally disordered and adopt a functional folding upon interaction with ubiquitin (Awakumov *et al*., 2006; Hu *et al*., 2005; Reyes-Turcu *et al*., 2008). In addition to ubiquitin, at least 16 other proteins encoded by the human genome exhibit characteristic ubiquitin fold collectively referred to as ubiquitin-like proteins for example SUMO protein, certain neuronal precursors and Interferon Stimulated Gene 15 (ISG15) (Hochstrasser, 2009). Upon interaction with the target protein, the ubiquitin molecule is identified by USPs from the unique stretch of 6 residues at the C-terminus of ubiquitin which differs from other ubiquitin-like molecules however, cross reactivity is not uncommon (Catic *et al*., 2007; Drag *et al*., 2008; Malakhov *et al*., 2002). Ubiquitin forms an isopeptide bond with its partner protein which differs from a conventional peptide bond on the basis of free rotation of bonds (Komander *et al*., 2009). USPs, are cysteine peptidases and exhibit isopeptidase activity to remove the attached ubiquitin. Cysteine dependent DUBs (like USPs) have a catalytic diad or triad of amino acids which mechanistically acts similar to the well studied plant cysteine peptidase, papain (Johnston *et al*., 1997; Storer and Menard, 1994). In short, the catalytic cysteine conducts a nucleophilic attack on the isopeptide bond and this requires lowering of the pKa of Cys, which is facilitated by a proximal polarized His residue. With few exceptions, polarization of His is further dependent on Asp or Asn alignment with the His. The catalysis is carried out by the hydrolysis of the acyl Cys intermediate which is formed by covalent association of the carboxyl group of ubiquitin with the enzyme.

Although the main function of USPs in the cell is to maintain adequate levels of the functionally important proteins and to recycle the pool of free ubiquitin, an increasing body of evidence demonstrates that they are directly and indirectly involved in variety of different biological functions. Given the diversity of USPs it is difficult to list let alone describe all the biological roles of all USPs, however, the functional distribution of USPs is illustrated (Figure 1.7) and summarised below.

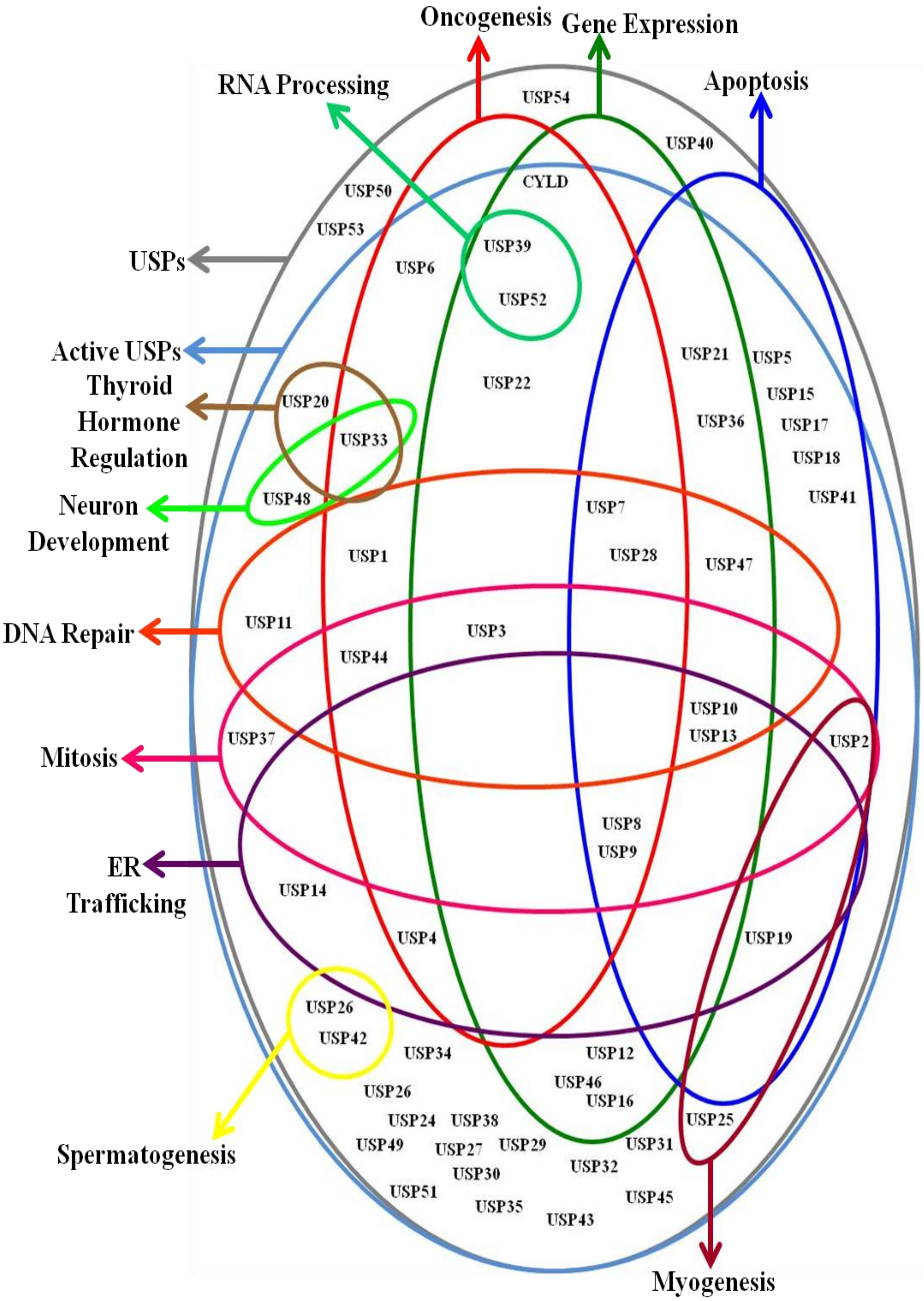


Figure 1.7. USPs biological roles: Venn diagram depicting the distribution of USPs according to their functions as reported in the literature. For the sake of clarity some minor or relatively less established functions are not represented. Important functions are shown here (differently coloured) encircling the involved USPs.

1.7.1. The role of USPs in gene expression

Eukaryotic genomic DNA is tightly wrapped around histones forming small nucleoprotein structures referred to as nucleosomes. Each nucleosome consists of a 146bp segment of DNA wrapped around an octamere of histone proteins. The tails of each histone molecule protrudes out of the nucleosome forming the target of several posttranslational modifications like methylation, acetylation, phosphorylation, sumolyation and ubiquitination (Strahl and Allis, 2000). These posttranslational modifications regulate chromatin structure which in turn regulate gene expression, DNA repair and chromosome condensation. Several E2 and E3 ubiquitin ligases are known to modify histones (Hammond-Martel *et al*., 2012). Similarly, several DUBs (including USPs) have been identified opposing the actions of the ubiquitin ligase on histones. For instance USP3, USP7, USP12, USP16, USP21, USP22, USP36 and USP46 deubiquitinate either H2A, or H2B, or both which variably results in repression or activation of genes (Joo *et al*., 2007; 2011; Nakagawa *et al*., 2008; Nicassio *et al*., 2007; Taillebourg *et al*., 2012; Zhang *et al*., 2008;). Other gene regulatory mechanisms affected by USPs include: preventing the degradation of cytoplasmic mRNA by USP52 (Bett *et al*., 2013), USP39 involvement in the RNA processing (Rios *et al*., 2011) and transcriptional negative feedback loop by USP8 (Luo *et al*., 2012).

1.7.2. The role of USPs in apoptosis

Several USPs have been reported to have a role in the molecular machinery of apoptosis. Most USPs are proapoptotic (USP8, USP10, USP15, USP28, USP47 and CYLD) however, some are anti apoptotic (USP2, USP9 and USP18) and some may have a dual role (USP7) (Ramakrishna *et al*., 2011). Proapoptotic activities of USPs are generally mediated by stabilizing proteins involved in programmed cell death, for example USP7 and USP10 deubiquitinate p53, a key pro apoptotic protein (Li *et al*., 2004; Yuan *et al*., 2010). USP8 deubiquitinates Nrdp1, a ubiquitin E3 ligase, responsible for the proteosomal degradation of apoptosis inhibitor, Baculoviral IAP Repeat Containing 6 (BIRC6), and consequently lead to the induction of apoptosis (Qiu *et al*., 2004). USP15 stabilizes procaspases-3 leading to its dissociation from Skp, Cullin, F-box containing (SCF) complex to cause apoptotic cell death (Xu *et al*., 2009). USP28 stabilizes ChK2 and 23 BP1 (DNA damage response proteins) which regulate p53 mediated induction of proapoptotic genes (Zhang *et al*., 2006). CYLD stabilizes RIP1 which is essential for NF-kB activation (Wang *et al*., 2008). The anti apoptotic property of USP2 is mediated by its ability to deubiquitinate fatty acid synthase (FAS) (Graner *et al*., 2004). Although the mechanistic details of how inhibition of FAS promote apoptosis is not clear, however, only it has been recently demonstrated that the FAS upregulates an oncogenic protein, β catenin (Gelebart *et al.*, 2012). A pro apoptotic functioning of USP17 has been demonstrated; however the mechanistic details are elusive (Shin *et al.*, 2006). USP7 stabilizes p53 and its ubiquitin ligase, MDM2, the proapoptotic role of USP7 is thought to be the result of tight regulation between these contrasting functions (Li *et al.*, 2004). USP9X mediates its anti apoptotic activity by stabilizing apoptosis signal regulating kinase 1 (ASK1) and MCL1. MCL1 is a member of BCL2 family and normally required to ensure the survival of stem cells (Noguchi *et al.*, 2008; van Delft *et al.*, 2006). Antiapoptotic function of USP18 has also been demonstrated recently (Potu *et al.*, 2010).

1.7.3. The role of USPs in cancers

Since cellular processes like DNA repair, mitosis and apoptosis are all affected through the molecular events of oncogenesis (Hoeijmakers, 2009; Hussain *et al.*, 2009; Singh *et al.*, 2010), the involvement of several associated USPs with cancers is not surprising. Indeed the Oncomine database (Rhodes *et al.*, 2004) shows that dysregulation of several USPs has been observed in different cancers. Though the molecular mechanisms of many of these associations are poorly understood, aberration in ubiquitin mediated degradation plays a central role in this regard. For example, premature truncation of CYLD translation due to germ line mutation (stop codon) has been associated with cylindromatosis and trichoepthelioma (Bignell *et al.*, 2000; Massoumi *et al.*, 2007; Poblete Gutierrez *et al.*, 2002). It has been demonstrated that CYLD is a negative regulator of the NF-kB signalling cascade, a pathway known for its oncogenic consequences (Baud *et al.*, 2009). Similarly, over expression of USP16 due to chromosomal translocation at chromosome 17p13 has been found to be causative in many aneurysmal bone cysts (Oliveira *et al.*, 2006). USP8 is another DUB implicated for its role in oncogenesis due to its role in regulating receptor tyrosine kinase (RTK) internalization and proteolytic degradation and consequently cell proliferation (McCullough *et al.*, 2004; Row *et al.*, 2006). USP9 stabilizes β-catenin and SMAD4 (important components of Wnt and TGFβ pathway respectively) suggesting its role in cellular proliferation and potentially oncogenesis (Dupont *et al.*, 2009; Murray *et al.*, 2004). USP3 and USP21 deubiquitinate histone subunits and the former may also be involved in DNA repair; both genes have been shown to be dysregulated in a variety of cancers (Oncomine database). As well as several other USPs (USP1, USP3 and USP28) are involved in DNA repair processes and thus their deregulation may be associated with cancer (Hussain *et al.*, 2009).

1.8. Chitinase and chitinase like proteins

Chitin, the linear polysaccharide of N-acetylglucosamine, is ranked second after cellulose in terms of abundance in nature. It serves as a structural component of many invertebrates including the exoskeleton of crustaceans and insects, shells and radulae of gastropods, the internal skeleton of cephalopods and the microfibrial sheet of parasitic nematodes. It is also a major constituent of fungal cell walls and in some cases it is found in structural elements of lower chordates and fishes such as *Branchiostoma floridae* and *Paralipophrys trigloides* respectively (Guerriero, 2012; Tharanathan and Kittur, 2003; Wagner *et al*., 1993; Weaver *et al*., 2011). Proteins required for the hydrolysis and/or remodelling of chitin are referred to as chitinases. They are found in all taxa of living organisms from bacteria to primates (Arakane and Muthukrishnan, 2010; Kasprzewska, 2003; Ohno *et al*., 1996).

Chitinases are members of the glycoside hydrolases (GH) protein family which is one of the largest and most diverse group of proteins. They are classified into 14 clans and 133 Carbohydrate-Active Enzyme database (CAZy) families on the basis of sequence and structural similarity, substrate specificity and catalytic mechanism (Cantarel *et al*., 2009). Chitinases are generally restricted to the GH18 and GH19 protein families and to a very limited extent are members of GH20 and GH48 families (Fujita *et al*., 2006; Kubota *et al*., 2004). Chitinases of the GH18 family carry out their catalytic function by substrate assisted mechanism while GH19 enzymes employs single displacement or inverting mechanisms for catalysis (Brameld and Goddard, 1998; van Aalten *et al*., 2001). The difference in catalytic mechanism and structural features suggests independent evolutionary lineages for members of these families. Broadly, chitinases can also be classified as endo or exo chitinases. Endochitinases perform internal but random cleavage of the polymer, while exochitinases mainly act on terminal ends of branched and unbranched polymers (Dahiya *et al*., 2006).

During the course of evolution, higher plants and vertebrates have replaced chitin by cellulose and hyaluronan respectively, yet plants and animals bear genes encoding active chitinases. Plant encoded chitinases are included in both GH18 and GH19 families of classes I, II and IV, where as animal encoded chitinases almost exclusively (except for some nematodes) are members of the GH18 family (Kasprzewska, 2003). All vertebrates (excluding some fishes) do not synthesize chitin and most species do not use chitin as a nutritional source. However, antifungal, antiprotozoal and antihelminthic properties have been attributed to human chitinases (Barone *et al*., 2003; Boot *et al*., 1998; 2001). Moreover, CHIA is found associated with the pathophysiology of asthma (Zhu *et al*., 2004) and others diseases involving immune dysfunctions (Lee, 2009; Sutherland *et al*., 2011). Therefore the presence of chitinases (particularly CHIT1 due to its elevated expression in macrophages in humans) is thought to be linked with immunity against chitin containing pathogens.

Humans have three catalytically active chitinases, two endochitinases: namely chitotriosidase I (CHIT1) and acidic mammalian chitinase (AMCase or CHIA), and an exochitinase chitobiase (CTBS). In addition to this, humans also encode four sequentially and/or structurally related inactive chitinases termed chilectins (ChiLs). These are CHIL1 (aka: CHI3L1, YKL-40 and CGP-39), CHIL2 (aka: CHI3L2, YKL-39, CP-39), oviductin (OVGP1) and stabilin 1 interacting chitinase like protein (CHID1) (Kzhyshkowska *et al*., 2006). These proteins lack catalytic activity due to substitution from a glutamic acid residue in the catalytic region of the protein but CHIL1 (at least) retains the ability to bind with chitin (Bussink *et al*., 2006; Houston *et al*., 2003). In mice, except for CHIL2, the other three ChiLs (CHIL1, OVGP1 and CHID1) are present. Intriguingly, an additional array of ChiLs has also been identified in mice: Chil3 (aka: Chi3l3, YM1), Chil4 (aka: Chi3l4, YM2), Chil5 (aka: Chi3l7, Bclp2) and Chil6 (aka: basic YM, BYm) (Hussain and Wilson, 2013). Phylogenetic analyses have shown that OVGP1 and all murine ChiLs are evolutionary related to CHIA, while CHIL1 and CHIL2 result from gene duplications of ancestral CHIT1 (Bussink *et al*., 2007; Funkhouser and Aronson, 2007).

Similar to active chitinases, studies have shown that most ChiLs are also involved in immunomodulation. For instance over expression of CHIL1 and in some cases also of CHIL2 has been reported in chronically inflamed tissues and in patients suffering from a variety of autoimmune disorders and cancers (reviewed in Coffman, 2008; Lee *et al*., 2011). Additionally, expression of ChiLs has also been found to be elevated in animal models of inflammatory diseases including allergy, asthma and cancer (Hannigan *et al*., 2011; Qureshi *et al*., 2011; Zhao *et al*., 2007). Importantly, up-regulation of ChiLs has been observed in EBV associated Hodgkin’s lymphoma, breast cancer, gastric carcinoma and nasopharyngeal carcinoma (Biggar *et al*., 2008; reviewed in Ober and Chup, 2009). ChiLs are also up-regulated in LMP1 (EBV oncogene) transgenic mice (Hannigan *et al*., 2007 Qureshi *et al*., 2011). Another chilectin, OVGP1 is normally expressed in the ovary and cervix and functions in fertilization and early embryo development (reviewed in Buhi *et al*., 2002; Lindsay *et al*., 1999; Yong *et al*., 2002). Like other ChiLs, elevated OVGP1 expression has been observed in the inflammatory disorders, such as disendometriosis (Wang *et al*., 2009) and ovarian cancer (Maines-Banidiera *et al*., 2010).

To date GH18 domain (39kDa) of two active chitinases of humans (CHIT1 and CHIA) and three human ChiLs CHIL1, CHIL2 and CHID1 have been structurally resolved by x-ray crystallography (Fusetti *et al*., 2002; Houston *et al*., 2003; Meng *et al*., 2010; Olland *et al*., 2009; Schimpl *et al*., 2012). In addition, structures of two other ChiLs, one from family *bovidae* termed BP40 and Chil3 from the mouse, have been resolved (Srivastava *et al*., 2007; Sun *et al*., 2001). Structurally, all active chitinases and ChiLs contain a 39KDa GH18 homology domain which is composed of two sub-domains, a large triosephosphate isomerase (TIM) β barrel domain and a relatively small α+β domain (Figure 1.8). Based on the presence or absence of the latter, GH18 family chitinases are further classified into subfamily A or B respectively. To date, all the known vertebrate GH18 homologues contain both the TIM barrel and the α+β domains and thus belong to family A. Distribution of family B chitinases is mostly restricted to bacteria (Suzuki *et al*., 2002; Watanabe *et al*., 1993). The TIM barrel domain is comprised of a β barrel like structure which is composed of eight anti parallel β strands and eight α helices. This $(\alpha/\beta)_8$ fold has been found as a primary structural component of many catalytic proteins, varying greatly in terms of sequence identity and function (Nagano *et al*., 2002). In tertiary conformation, this TIM barrel (with little contribution from the α+β domain) forms a ligand binding cleft lined with the solvent exposed aromatic residues. The catalytic site lies within this cleft of active chitinases, characterized by a DxDxE motif where Asp and Glu are required for the catalytic activity (Fusetti *et al*., 2002; Olland *et al*., 2009). These amino acids are replaced by Leu/Ile or Gln in different ChiLs (Sun *et al*., 2001; Houston *et al*., 2003). Recently, on the basis of primary sequence conservation, another sub domain has been identified as a chitinase insertion domain (CID), located between the 7th α helix and 7th β strand of the TIM barrel. Computational analyses have shown the importance of this domain in ligand binding affinity and specificity (Li *et al*., 2010). In addition to the GH18 domain, another relatively small chitin binding domain, CBM14, has also been described at the C-terminal end of both active chitinases CHIT1 and CHIA. This domain is also present in many invertebrate chitinases and the CBM14 of CHTI1 has been shown to bind with insoluble (colloidal) chitin (Tjoelker *et al*., 2000). The C-terminal region of human OVGP1 instead consists of a long tail with patchy similarity to mucin-like proteins. The tail has no recognizable sequence similarity with CBM14 and predicted to be heavily glycosylated (Huang *et al*., 2012). Taken together, we have proposed that the phylogenetic relationship and subtle variations in the structures of vertebrate chitinase and ChiLs are the combined product of shared ancestry and independent evolution leading to their structural and functional divergence (Hussain and Wilson, 2013).

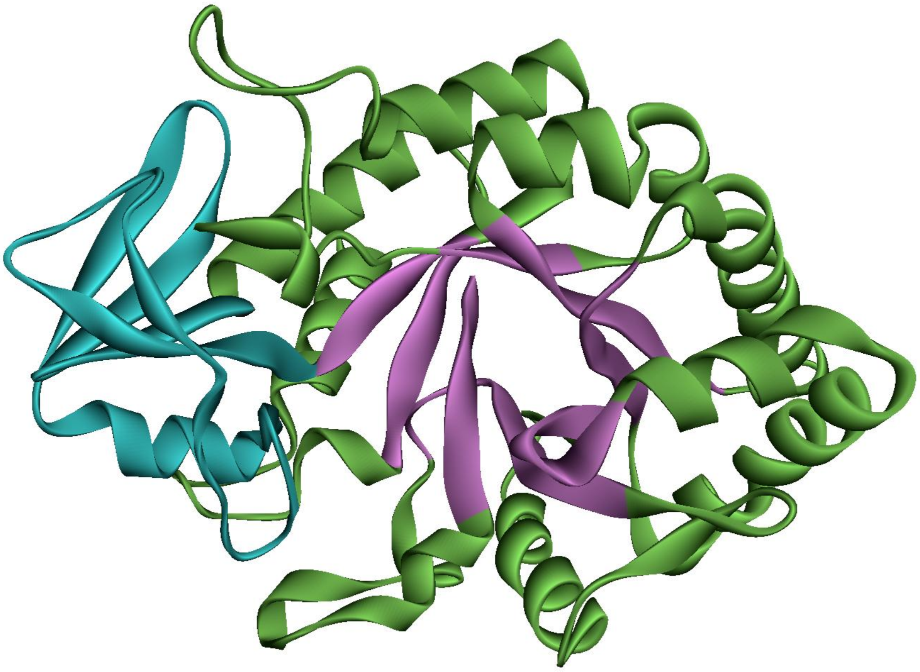


Figure 1.8. Structure of CHIL1. Ribbon structure of human CHIL1 (PDBid: INWU) is shown here with important functional regions indicated: TIM barrel domain (green), α+β domain (cyan) and ligand binding central cleft (pink).

1.9. Aims and approaches of project

1.9.1. Chapter 3. EBNA1 structure-function relationship

The purpose of these studies was to develop and analyse a full length model of EBNA1 of human and other primate LCVs using *in silico* methodologies. Upon generating a plausible model, the full length structural conformation of an EBNA1 dimer of human and other primates LCVs was constructed. A further aim was to design and screen an inhibitory peptide using *in silico* molecular docking approach. Preliminary tests, to examine the efficiency of inhibitory peptide, were conducted using cell culture based assay.

1.9.2. Chapter 4. EBNA1 protein-protein interactions

The main aims of the chapter were to explore the utility and efficiency of peptide arrays for investigating the protein interactions of EBNA1. Modified peptide (alanine replacement and truncated peptide) arrays were used to explore the interaction sites at fine resolution.

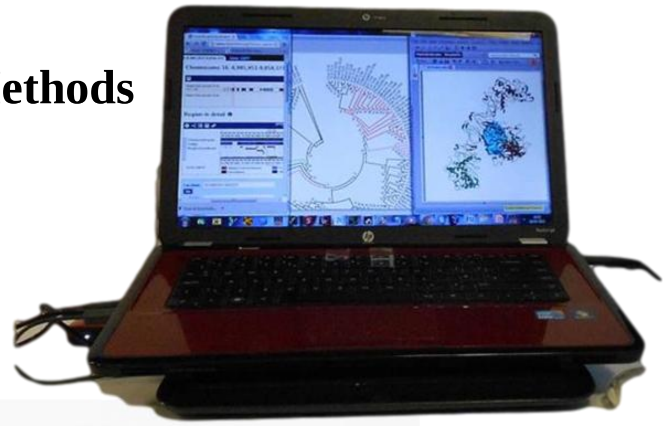
1.9.3. Chapter 5. Evolution of Ubiquitin Specific Proteases

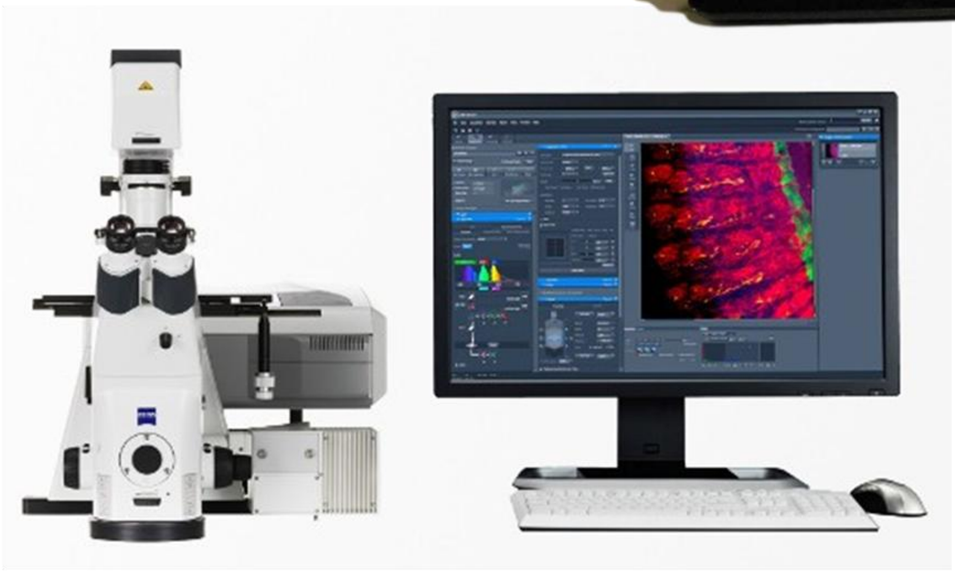
The aim of this work was to conduct extensive phylogenomic analyses of USP homologues in order to explore, the evolutionary time line of the diversification of USPs in animal kingdom. USPs homologues from animals representing major evolutionary lineages were selected for the study. Special attention was paid to determine the origin, expansion and functional divergence of USPs in animals in general and vertebrates in particular. Genomic synteny analyses were incorporated into the study to explore the different mechanisms at work in the evolution of USPs. Any unresolved phylogenies of the homologues were assessed using evolutionary distance analysis. Expression patterns and protein interaction networks were compared to assess functional innovations paralleling USP expansion.

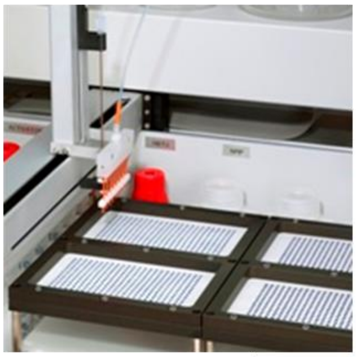
1.9.4. Chapter 6. Phylogenomic studies of chitinase and chitinase like proteins

The aim of this study was to determine the evolutionary history of the vertebrate Chitinases/ChiLs. A robust data mining of vertebrates GH18 family genes was carried out to examine the distribution of GH18 family members across different vertebrate lineages. Phylogenetic analyses were carried out to explore the evolutionary relationship of different GH18. Genomic synteny of GH18 family genes of selected vertebrates was compared to investigate the underlying mechanism of the expansion of GH18 family. Structurally unresolved GH18 family proteins (to complete the set of encoded proteins from paralogous genes) were modelled and assessed for their structural plausibility. Several structural features including ligand binding regions of the known and modelled structures of GH18 proteins were compared to explore the structural and functional divergence of the proteins.

Chapter 2. Materials and Methods







2.Materials and Methods

2.1. Buffers and reagents

Unless mentioned otherwise, all buffers were made in dH_2O

2.1.1. General buffers

PBS (Phosphate buffered saline)

80mM Na_2HPO_4

20mM NaH_2PO_4

100mM NaCl

PBST (PBS with Tween 20)

80mM Na_2HPO_4

20mM NaH_2PO_4

100mM NaCl

0.05% (v/v) Tween-20

TBST (Tris-phosphate buffered saline with Tween 20)

50mM Tris

pH to 7.5 with HCl

150mM NaCl

0.05% (v/v) Tween-20

2.1.2. Buffers and reagents used in nucleic acid procedures

TE

10mM Tris

1mM EDTA

pH to 7.5 with HCl

TAE

40mM Tris

2mM EDTA

20mM acetic acid

10x DNA loading dye

50% (v/v) glycerol

50% (v/v) TE pH 7.2

Trace of bromophenol blue

Trace of xylene cyanol

DEPC water

0.1% (v/v) di ethyl pyrocarbonate in dH_2O,

Shaken vigorously and left over night,

Autoclaved next day

4M GT stock

4M guanidinium thiocynate

25mM sodium citrate pH7.0

0.5% sarcosyl

dissolved at 65°C in DEPC water

MOPS

20mM MOPS

1mM EDTA

5mM sodium acetate

pH to 7.0 sodium hydroxide

Solution D

0.7% (w/v) β-mercaptoethanol in 4M GT stock

RNA loading dye

50% (w/v) glycerol

0.1% (w/v) bromophenol blue

0.1% (v/v) xylene cyanol

50% (v/v) TE pH 7.5

RNA loading buffer

1xMOPS

17.8% (v/v) formaldehyde

50% (v/v) formamide

Transformation Buffer 1:

30mM potassium acetate

100mM $RbCl_2$

10mM $CaCl_2.2H_2O$

50mM $MnCl_2.4H_2O$

15% (v/v) glycerol

Set pH to 5.8 using 0.2 M acetic acid

Volume made up to 200ml

Filtered to sterile

Transformation Buffer 2:

10mM MOPS

10mM $RbCl_2$

75mM $CaCl_2.2H_2O$

15% (v/v) glycerol

pH set to 6.5 with 0.2 M KOH

Volume made up to 100ml

Filtered to sterile

2.1.3. Buffers and reagents used in protein related procedure

Lysis buffer (for bacteria)

360mM NaCl

5mM imidazole (not included for bacteria with pGEX-6P-1 plasmid)

Dissolved in 1xPBS

Complete protease inhibitor 1 tablet/10 ml (freshly added)

Aprotinin 100µl/10 ml (freshly added)

1mM DTT (freshly added)

1mg/ml lysozyme* (freshly added)

*(only added for lysis)

Wash buffer-I (protein purification)

360mM NaCl

5mM imidazole (not included for bacteria with pGEX-6P-1 plasmid)

Dissolved in 1xPBS

set pH to 7.5

Elution buffer (His tag protein purification)

360mM NaCl

250mM imidazole*

Dissolved in 1xPBS

set pH to 7.5

Complete protease inhibitor 1 tablet/10 ml (freshly added)

Aprotinin 100µl/10 ml (freshly added)

*For gradient purification in addition to 250mM imidazole, buffer containing 50mM, 100mM, 150mM, 200mM and 300mM of imidazole were also prepared.

Elution buffer (GST tag protein purification)

10mM glutathione

Dissolved in 1xPBS

Wash buffer-II (protein purification)

23 parts wash buffer-I

2 parts elution buffer for His tag protein

RIPA buffer

150mM NaCl

50mM tris-HCl pH7.5

1% (v/v) triton X

1% (w/v) deoxycholic acid

0.1% SDS

Aprotinin 100µ1/10ml (freshly added)

Roche complete mini (protease inhibitor) 1 tablet/10ml (freshly added)

Protein loading dye

7.5% (v/v) glycerol

2.5% (v/v) β-mercaptoethanol

2% (w/v) SDS

50mM tris-HCl pH 6.8

bromophenol blue (trace)

Transfer buffer

25mM tris

92mM glycine

pH to 8.3 with HCl

20% (v/v) methanol

Running buffer

25mM tris

190mM glycine

0.1% (w/v) SDS

Stripping buffer

100mM β-mercaptoethanol

2% (w/v) SDS

62.5mM tris-HCl pH6.8

Fixing solution

50% (v/v) Methanol

10%(v/v) Acetic acid

Staining solution

50% (v/v) Methanol

10% (v/v) Acetic acid

0.05% (w/v) Comassie brilliant blue

Destaining solution

30% (v/v) Methanol

10% (v/v) Acetic acid

Blocking solution (western blot membrane)

5% non fat milk (NFM) in 1xPBST

Blocking solution (glass peptide array)

5% Bovine Serum Albumin (BSA) in 1xTBST

Blocking solution (membrane peptide array)

5% NFM in 1xPBST

Probing solution (glass peptide array)

0.5% Bovine Serum Albumin (BSA) in 1xTBST

Probing solution (membrane peptide array)

1% NFM in 1xPBST

Stripping solution (membrane peptide array)

60mM Tris

set pH to 6.8 with HCl

20mM DTT

2% (w/v) SDS

2.2. Chemical and reagents

Table 2.1. Table of the chemicals used

Chemicals	Company	Catalogue No.
1kb DNA ladder	Invitrogen	10787-01-8
2-mercaptoethanol	SIGMA-ALDRICH	6132-04-3
40% bis/acrylamide gel solution	BIORAD	161-0148
Acetic acid	BDH Labs	10001CU
Agarose	Invitrogen	16500-500
Ammonium per sulphate (APS)	SIGMA-ALDRICH	A3678
Ampicillin	SIGMA-ALDRICH	A9518
Aprotinin	SIGMA-ALDRICH	A6012
Biorad dye	BIORAD	500-000-6
Bovine serum albumin (BSA)	SIGMA-ALDRICH	A3311
Chloramphenicol	SIGMA-ALDRICH	C0378
Cobalt Slurry	Thermo Scientific	89965
Coomassie blue	Koch Light Labs	88755
DAPI	SIGMA-ALDRICH	D9542
Deoxycholic acid	SIGMA-ALDRICH	D6750
dNTPs	Invitrogen	10297-018
DTT	SIGMA-ALDRICH	D9779
ECL detection kit	GE health care	RPN2209
EDTA	SIGMA-ALDRICH	27285
Glutathione sepharose fast flow	GE health care	17-5132-01
Glycerol	Fisher-Scientific	BP229-1
Glycine	Fisher-Scientific	56-40-6
Guanidium thiocynate	SIGMA-ALDRICH	G9277
Imidazole	SIGMA-ALDRICH	15513
IPTG	Melford	MB1008
Isopropanol	SIGMA-ALDRICH	24137
Kanamycin	SIGMA-ALDRICH	K4378
Methanol	Fisher Chemical	M/4000/PC17
Miniprep plasmid extraction kit	Qiagen	27104
Midiprep plasmid extraction kit	Qiagen	12143
NaCl	VWR International	27810-295
Na_2HPO_4	VWR International	102494C
NaH_2PO_4	Riedel-deHaen	04269
NaOH	Fisher Chemical	S/4920/53
Nickel Slurry	Qiagen	1018244
Oligo(dT)$_{20}$ Primer	Invitrogen	18418-020
Phosphatase inhibitor	Roche	04906837001
Precision Plus Protein ladder	BIORAD	161-0373
Protease inhibitor	Roche	11873580001
QIA quick gel extraction kit	Qiagen	28104
QIA quick PCR purification kit	Qiagen	28704
Sarcosyl	SIGMA-ALDRICH	L5125
SDS	VWR International	442444H
Sodium citrate	Fisher Scientific	6132-04-3
TEMED	SIGMA-ALDRICH	T9281
Tris	Fisher Scientific	BP152-1
Triton	SIGMA-ALDRICH	57H0650
Tween-20	SIGMA-ALDRICH	P5927

2.3. Cell culture

Table 2.2. Table of reagents used for cell culture

Chemicals	**Company**	**Catalogue No.**
Fetal Bovine Serum (FBS)	SIGMA-ALDRICH	F9665
L-Glutamine (2% v/v of 200mM stock)	SIGMA-ALDRICH	G7513
PBS	GIBCO	14190-094
Penicillin-Streptomycin solution (2% v/v of 200mM Stock)	Lonza	17603E
RPMI-1640	SIGMA-ALDRICH	R0883
Trypan blue (0.4% solution)	SIGMA-ALDRICH	T8154

2.4. Enzymes

Table 2.3. Table of the enzymes used

Enzymes	**Company**	**Catalogue No.**
*BamH*I	New England Biolab	R3136S
*EcoR*I	New England Biolab	R3101S
*Hind*III	New England Biolab	R3104S
MMLV-Reverse transcriptase	Invitrogen	14190-094
Phusion high fidelity polymerase	New England Biolab	M0530S
T4 DNA ligase	Invitrogen	15224-017
T4 DNA polymerase	Invitrogen	18005-017
RNase free DNAase	Promega	M6101
*Xho*I	New England Biolab	R0146S

2.5. Antibodies

Table 2.4. Table of the antibodies used

Antibodies	**Dilutions**	**Supplier**	**Cat.No**
IH4	1:20	Snudden *et al*., 1994	---
Aza2E8	1:200	Hearing *et al*., 1985	---
Rab 16-4	1:1000	Prof. J.M. Middeldorp	---
EBP2	1:1000	Santacurz	sc-46314
USP7	1:1000	Abcam	ab4080
Anti-goat-HRP	1:4000	Santacurz	sc-2032
Anti-mouse-HRP	1:4000	Santacurz	sc-2031
Anti-rabbit-HRP	1:4000	Santacurz	sc-2030
Anti-rat-HRP	1:4000	Santacurz	sc-2032
Anti 6x-His-HRP	1:4000 (western) 1:6000 (slide peptide array) 1:5000 (membrane peptide array)	Abcam	ab1187

2.6. Cell lines

Table 2.5. Table of the cell lines used

Cell lines	EBV status
BJAB	EBV negative
B958	EBV positive
BL2+ve	EBV positive
BL2-ve	EBV negative
BL30+ve	EBV positive
BL30-ve	EBV negative
BL70+ve	EBV positive
BL70-ve	EBV negative
IB4	EBV positive
Namalwa	EBV positive
Raji	EBV positive

2.7. Primers

All primers were purchased from Sigma Aldrich and presented in following tables

Table 2.6. Table of the primers used

Primers	Direction	Sequence
EBP2FG	Forward	GATCTTGCTCGAGATGCTATTTAGTGTGTTCTG
EBP2RG	Reverse	CTGACGGATCCGGCGAGATGGACACTCC
EBP2FH	Forward	GTGCATAAGCTTGGCCACCATGGACACTCCCCCG
EBP2RH	Reverse	GATCTGCTCGAGATGCTATTTAGTGTGTTCTG
USP7F	Forward	GTGCATAAGCTTGATGAACCACCAGCAG
USP7R	Reverse	GTGTGTCATATCTCGAGCAGCTTGGAAATCAGTTATG

2.8. Plasmids

2.8.1. pGEX-6P-1

pGEX-6P-1 (GE Healthcare, 28-9546-48) is a 4.9Kbp bacterial expression plasmid which tags Glutathione S-Transferase to the N-terminal of the recombinant protein. The plasmid harbours ampicillin resistance gene for selection and the inserted gene expression is induced by isopropyl β D thiogalactopyranoside (IPTG) under tac promoter. *Bam*HI and *Xho*I sites were used for cloning human EBP2 in the plasmid.

2.8.2. pET-28c

pET-28c (Novagen 69866-3) is a 5.3Kbp bacterial expression plasmid which tags 6-histidine residues to the N-terminus of the recombinant protein. The plasmid harbours kanamycin resistance gene for selection and the inserted gene expression is induced by IPTG under T7 promoter. *Hind*III and *Xho*I sites were used for cloning human EBP2 and USP7 in the plasmid.

2.9. Peptides

Dimerisation Inhibiting Peptides (DIP) designed and used in this study were synthesized and purchased from Cambridge Peptides.

Table 2.7. Sequences of the peptides used

Peptides	Sequences
DIP	YGRKKRRQRRRFGMAPGPGPQPGPLR
DIP-Flu	YGRKKRRQRRRFGMAPGPGPQPGPLR-Flu*

*Flourescein molecule is attached to the C-terminal of the peptide for microscopic studies

2.10. Bacterial Strains

2.10.1. *Escherichia coli* DH5α

E.coli DH5α is an efficient bacterial strain for cloning, primarily because of the deletion in restriction endonucleases *endA1* and *hsdR17* genes.

2.10.2. *Escherichia coli* BL21 star (DE3)

Competent *E.coli* BL21 was purchased from Invitrogen (Cat#440049). The strain offers high mRNA and protein stability due to mutation in *rne131, lon* and *OmpT* genes.

2.10.3. *Escherichia coli* Rosetta 2(DE3) plysS

Competent *E.coli* Rosetta 2 was purchased from Novagen (Cat#714013). The strain is a derivative of BL21 strain with an additional plasmid encoding seven rare eukaryotic tRNA.

2.11. Bacterial media

2.11.1. LB broth

1% (w/v) tryptone

0.5% (w/v) yeast extract

1% (w/v) NaCl

pH set to 7.5 by NaOH

2.11.2. LB agar

1% (w/v) Tryptone

0.5% (w/v) Yeast extract

1% (w/v) NaCl

pH set to 7.5 by NaOH

1.5% (w/v) Agar

2.12. Computation

2.12.1. HP Pavilion g6 notebook PC

Intel (R) Core (TM) i5 CPU

M480 @2.67 Ghz

RAM 4.00 GB

Windows 7 64 bit

Linux red hat (virtual)

2.12.2. Mac OS X Version 10.6.2

Intel Core 2 Duo

3.06 Ghz

RAM 8.00 GB

Mac OS X Version 10.6.2

2.12.3. Servers and Softwares

Table 2.8. Table of softwares and server used

Programmes /Server	**Main purpose**	**References**
Swiss model	Homology modelling	Kiefer *et al*., 2009
Modeller	Homology modelling	Eswar *et al*., 2006
ITASSER	Iterative multiple threading modelling	Roy *et al*., 2010
MOE	Homology and *ab initio* modelling	ChemComp
ClusPro	Molecular Docking	Comeau *et al*., 2004
SymmDock	Molecular Docking	Schneidman-Duhovny *et al*., 2005
RCSB data base	Protein structure retrieval	Berman *et al*., 2000
Fold Index	Structure propensity estimation	Prilusky *et al*., 2005
MolProbity	Structure assessment	Chen *et al*., 2010
Q mean square	Structure assessment	Benkert *et al*., 2009
SPDB Viewer v4.0.2	Structure assessment	Johanson *et al*., 2012
DS visualizer	Protein structure visualization	Accelrys
POCASA 1.0	Cavity analysis	Yu *et al*., 2010
NCBI server	Data mining and genomic synteny	Wheeler *et al*., 2008
Ensembl server	Data mining and genomic synteny	Flicek *et al*., 2013
HMMER	Data mining	Finn *et al*., 2011
UniProt server	Data mining and domain identification	Margrane *et al*., 2011
CDD server	Domain identification	Marchler-Bauer et al., 2011
BioGPS	Gene expression data	Wu *et al*., 2009
STRING 9.05	Protein interaction	Szklarczyk *et al*., 2011
Clustal X	Multiple sequence alignment	Thompson *et al*., 1997
Bioedit v 7.0.53	Multiple sequence alignment	Hall, 1999
CLC seq.viewer v5.1	Multiple sequence alignment visualization	CLC bio
NetNGlyc 4.0	Prediction of N glycosylation sites	www.cbs.dtu.dk
NetOGlyc 4.0	Prediction of O glycosylation sites	www.cbs.dtu.dk
MEGA5.1	Phylogenetic analysis	Tamura *et al*., 2011
Fig tree	Tree visualization	tree.bio.ed.ac.uk
Time tree	Timeline for speciation events	Hedges *et al*., 2006
pDraw32.1.0	Plasmid drawing	ACAclone
Prismv4.0.C	Graphs and statistical analysis	GraphPad

2.13. DNA and RNA techniques

2.13.1. RNA isolation

RNA was isolated from Raji cell line according to the method by Chomczynski and Sacchi (1987). The cell pellet was suspended in 250µl of solution D by vortexing and pipetting. 50µl of 2M sodium acetate was dispensed in the tube and after gentle vortexing, 500µl of phenol and 100µl of chloroform: isoamylalcohol (49:1) was added in the tube. The extract was mixed by inversion and vortexed for 10 seconds and was incubated on ice for 15 minutes. The samples were centrifuged at 10,000g for 20 minutes at 4^{o}C and the upper aqueous phase was gently transferred to a fresh tube. 500µl of prechilled isopropanol was added into extract and stored at -20^{o}C over night to precipitate the RNA. Next day RNA was pelleted by centrifugation at 10,000g for 20 minutes at 4^{o}C and the supernatant was removed. The pellet was resuspended in 300µl of solution D and 600µl of ethanol and stored at -20^{o}C for 2 hours. The precipitated RNA was recollected by centrifugation at 10,000g for 20 minutes at 4^{o}C and the supernatant was discarded. The pellet was washed twice with 75% ethanol by centrifuging for 5 minutes at 10,000g between each wash. The supernatant was removed and pellet was allowed to air dry before dissolving in 100µl of DEPC water. 1µl of RNA sample was used to quantify the concentration of extracted RNA using nanodrop at 260nm and 280nm. 5µl of the sample was mixed with 15µl of RNA loading dye and heated at 68^{o}C for 10 minutes. Finally, the samples were loaded in wells of 1% agarose gel (prepared in 1xTAE in DEPC water) to evaluate the quality of RNA. Sharp bands representing 28S, 18S and 5S rRNA reflect the presence of good quality RNA.

2.13.2. RNA purification

DNA contamination from RNA was removed using RQ1 RNase free DNase. 4µl of DNase, 6µl of 10x buffer was mixed and DEPC water was added to a total volume of 60µl. The reaction mixture was incubated for 1 hour at 37^{o}C and subsequently the enzyme was inactivated by heating the mixture to 75^{o}C for 10 minutes. DEPC water was added to make the total volume of 250µl, after which an equal volume (250µl) of phenol solution saturated with 0.1M citrate buffer was dispensed and mixed by inversion. The mixture was centrifuged at 14,000g for 2 minutes at room temperature and upper aqueous phase was removed and transferred to a new tube. 250µl of chloroform was added and mixed by inversion. The sample was centrifuged at 14,000g for 2 minutes at room temperature and the aqueous phase was transferred to a fresh tube. 625µl of ethanol and 24µl of 3M sodium acetate was dispensed and the mixture was stored at -20^{o}C over night for precipitation. The RNA was pelleted by centrifugation at 14,000g for 30 minutes at 4^{o}C and the supernatant was removed. The pellet was washed with 1ml of 75% ethanol followed by centrifugation at 14,000g for 5minutes at 4^{o}C. After discarding the supernatant, the pellet was airdried briefly and dissolved in 20µl of DEPC water and stored at -20^{o}C.

2.13.3. cDNA synthesis by Reverse Transcription (RT-PCR)

The reaction mixture was prepared by mixing 5µg of DNase treated RNA, 1µl of oligodT and 6µl of DEPC water. The mixture was heated at 65^{o}C for 5 minutes to denature any secondary structure of RNA and to allow annealing of oligodT to the mRNA. 4µl of 5x reaction buffer, 2µl of dNTP (20mM) and 1µl of reverse transcriptase was added to the positive reverse transcription (RT+ve) reaction mixture. To monitor the possibility of DNA contamination, a RT-ve reaction mixture was also prepared with similar composition except 1µl of reverse transcriptase was replaced by 1µl of DEPC water. The reaction mixture was incubated at 37^{o}C for 1 hour and then enzyme was inactivated by heating at 75^{o}C for 15 minutes. This reverse transcribed cDNA mixture was then used to amplify the specific product by conventional PCR.

2.13.4. Plasmid Isolation

Miniprep Plasmid Isolation: Miniprep plasmid isolation was carried out using Qiagen miniprep plasmid extraction kit for verifying the clones (inserts) by restriction digest. 5ml overnight culture of bacteria containing cloned or vector plasmid was pelleted by centrifugation at 10,000g for 5 minutes. The bacterial pellet was suspended in 250µl of suspension buffer (P1) by rigorous vortexing. 250µl of lysis buffer (P2) was dispensed in the tube and was left at room temperature for 5 minutes after mixing by inversion. 350µl of neutralizing buffer (N3) was added, mixed by inversion and tube was centrifuged at 15,000g for 10 minutes. The supernatant was transferred to a QIAprep spin column and centrifuged for 1 minute. The flow through was discarded and 500µl of binding buffer (PB) was dispensed over the bottom end of spin column before centrifugation for 1 minute. Flow through was discarded and column was washed by 750µl of wash buffer containing ethanol (PE). Flow through was discarded and tubes were centrifuged for one minute to remove any residual ethanol. The collection tube was replaced by a fresh eppendorf tube and 40µl of elution buffer (EB) was dispensed over the bottom end of column and centrifuged at 10,000g for one minute. The eluted plasmid was stored at -20^{o}C till further use.

Midiprep Plasmid Isolation: Qiagen plasmid midiprep protocol was used to obtain sufficient yield of the plasmids (require for cloning and/or sequencing). 100ml of the overnight culture of bacteria containing cloned or vector plasmid was pelleted by 10,000g for 15 minutes at 4^{o}C. The bacterial pellet was suspended in 4ml of suspension buffer (P1) by rigorous vortexing. 4ml of lysis buffer (P2) was dispensed in the tube and was left at room temperature for 5 minutes after mixing by inversion. 4ml of pre chilled neutralization buffer (P3) was added in the mixture and tube was incubated for 15 minutes on ice. The tube was centrifuged at 15,000g for 30 minutes at 4^{o}C and the supernatant was transferred to fresh tube immediately and centrifuged again at 15,000g for 20 minutes at 4^{o}C to remove any residual bacterial debris. QIAGEN-tip100 column was equilibrated by 4ml of equilibration buffer (QBT), passed by gravity flow. The supernatant was decanted in the pre-equilibrated column and allowed to pass through gravity flow. Washed the column twice with 10ml washing buffer (QC) and DNA was eluted by passing 5ml of elution buffer (QF) through the column. 3.5ml of isopropanol was added in the final flow through and centrifuged for 30 minutes at 4^{o}C. The pelleted DNA was washed twice by 5ml of 70% ethanol and after centrifugation the supernatant was discarded gently. The pellet was left for 10-15 minutes to air dry and DNA is dissolved in 1-1.5ml of 1xTE.

2.13.5. Polymerase Chain Reaction (PCR)

Target genes (EBP2 and USP7) were amplified using conventional PCR. The reaction mixture include: 5µl of reverse transcribed cDNA (EBP2) or 1µl (40-50ng) plasmid (pGEX-6P-1 EBP2 or pCI-neo Flag HAUSP (addgene 16655)) were mixed with 0.2µl of Phusion high fidelity polymerase, 4µl 5x polymerase buffer, 1µl dNTP (20mM), 1µl each of reverse and forward primers (10µM) and appropriate volume of autoclaved and filter sterile water to a total volume of 20µl. The PCR reaction was carried out for both RT+ve and RT-ve and plasmid DNA samples under following conditions.

1. Initial denaturation at 95^{o}C for 10 minutes
2. Denaturation at 95^{o}C for 5 minutes
3. Annealing at 66^{o}C for 5 minutes
4. Amplification at 72^{o}C for 3 minutes
5. Final amplification/extension at 72^{o}C for 10 minutes
6. Cycles: From step 2 to step 4, 30 cycles

All PCR reactions were loaded in 1% agarose gel and run in 1xTAE buffer and visualized for products (EBP2;1Kbp, USP7;3.3Kbp) under UV.

2.13.6. Purification of PCR products

For cloning, PCR products of the gene of interest (EBP2 and USP7) were cleaned using Qiaquick PCR purification kit. 20µl of PCR product was mixed with 100µl of buffer PB and dispensed on the Qiaquick spin column. The column was centrifuged at 10,000g for 1 minute and flow through was discarded. The column was then washed with 750µl of wash buffer (PE) and after centrifugation for 1 minute the flow through was discarded. The tubes were centrifuged again to remove any residual ethanol (present in wash buffer) and the DNA was eluted by dispensing 40µl of the elution buffer (EB) over the bottom of spin column. The tube was centrifuged for 1 minute at 10,000g and the eluted DNA was stored at -20^{o}C till further use.

2.13.7. Preparation of competent cells

A loop full glycerol stock of *E.coli* DH5α was inoculated in 20ml LB broth and incubated at 37^{o}C over night in shaking conditions (200rpm). 100µl of the overnight culture was dispensed in 5ml of LB broth and incubated at 37^{o}C under shaking (200rpm) till the OD reached at 0.3 at 550nm. Whole culture was transferred to 100ml of LB broth and incubated under the same condition till the OD reached between 0.4-0.5 at 550nm. The bacterial culture was divided into two equal halves and placed on ice for 5 minutes before centrifuging at 3000g for 5 minutes at 4^{o}C. Supernatant was discarded and each bacterial pellet was suspended in 20ml of transformation buffer-I and placed on ice for 5 minutes. The cell suspension was centrifuged (3000g) for 5 minutes at 4^{o}C and supernatant was discarded. Each bacterial pellet was resuspended in 5ml transformation buffer-II and 200µl of suspension was dispensed in separate pre autoclaved fresh tubes and immediately dropped into liquid nitrogen for snap freezing. The cells were stored at -80^{o}C till further use.

2.13.8. Cloning of human EBP2 and USP7 genes in expression vector

Both cleaned PCR products and plasmid were digested to create sticky end for cloning. EBP2 was cloned both in pGEX-6P-1 and pET-28c expression vectors while USP7 was only cloned in pET-28c vector. Following reaction mixtures were used for digesting different plasmids and PCR products.

Table 2.9. Reaction mixture for restriction digests

Ingredients	**For cloning in pGEX-6P-1**		**For cloning in pET-28c**		
	EBP2	**pGEX-6P-1**	**EBP2**	**USP7**	**pET-28c**
DNA	26µl	11µl	26µl	26µl	11µl
BamHI	2µl	2µl	--	--	--
XhoI	2µl	2µl	2µl	2µl	2µl
HindIII	--	--	2µl	2µl	2µl
10xbuffer	5µl	3µl	5µl	5µl	3µl
dH2O	15µl	12µl	15µl	15µl	12µl
Total	50µl	30µl	50µl	50µl	30µl
	Incubated at 37^{o}C over night				
*BamH*1	1µl	1µl	--	--	--
*Xho*I	1µl	1µl	1µl	1µl	1µl
*Hind*III	--	--	1µl	1µl	1µl

All mixtures were incubated at 37^{o}C for overnight and on next day 1µl of respective restriction enzymes was added and incubated for further 2 hours at 37^{o}C. Enzymes were then deactivated by heating at 65^{o}C for 20 minutes. Appropriate amount of DNA loading dye was added into mixture and samples were electrophoresed in 1% agarose in 1xTAE. The desired size bands were cut and access gel was removed. DNA was extracted from the gel by QIAquick gel extraction kit. Briefly, one volume of gel (1gm~1ml) was mixed with 3 volume of extraction buffer (QG) and sample was incubated at 50^{o}C for 10 minutes on thermomixer at shaking. One gel volume of isopropanol was added in the tubes, mixed by inversion and decanted in QIAquick spin column. The column was centrifuged at 10,000g for 1 minute and flow through was discarded. 500µl of the QG buffer was added to the column and centrifuged at 10,000g for 1 minute. The flow through was discarded and 750µl of wash buffer (PE) was dispensed in the column and centrifuged at 10,000g for 1 minute. The filterate was discarded and column was centrifuged again for 1 minute to remove residual ethanol (present in PE). Finally the column was fixed on a fresh eppendorf tube and 30µl of elution buffer was added at the bottom of column and tube was centrifuged at 10,000g for 1 minute. 1µl of the eluted DNA was used to quantify the amount of DNA by nanodrop and remaining was stored at -20^{o}C till further use.

Finally the gel extract, restricted plasmid and inserts were mixed at molar ratio 1:3 respectively by calculating the amount of the DNA required for insert against fixed amount of plasmid (vector) using following formula.

$$\text{ng of insert DNA required} = \frac{\text{ng of vector X Kbp size of insert}}{\text{Kbp size of vector}} \text{X} \frac{\text{molar ratio of insert}}{\text{molar ratio of vector}}$$

Appropriate volume of insert and plasmid DNA was mixed with 1µl DNA ligase and 5µl of 5x ligation buffer to a total volume of 20µl. The reaction mixture was incubated at 16^{o}C over night. Note; in the negative control insert DNA volume was replaced with equal volume of dH_2O. Next day the ligation mixture was used to transform competent cells.

2.13.9. Transformation of plasmids in bacteria.

Competent cells (*E.coli* DH5α or *E.coli* BL21 or *E.coli* Rosetta 2) were thawed and 10ng of DNA (from ligation mixture) was added in the tube. The bacterial DNA mixture was left on ice for 30 minutes to allow DNA to stick to the surface of the bacteria. After incubation the tubes were gently placed in the water bath set at 42^{o}C for 90 seconds after which tubes were promptly placed on ice for 1 minute. 200µl of prewarmed (37^{o}C) LB broth was added to the bacterial suspension and incubated at 37^{o}C on shaking for one hour. 100µl of the culture was inoculated on selective LB agar containing 50µg/ml ampicillin (for pGEX-6P-1 selection) or 50µg/ml kanamycin (for pET-28c selection) of 50µg/ml each of chloramphenicol and kanamycin (for Rosetta strain transformed with pET-28c plasmid). The inoculum was homogenously spread over the surface of plate and the plate was incubated at 37^{o}C for overnight.

2.13.10. Validation of cloning by colony PCR

A small part of colony of a clone was picked from the selection media and added into master mixture (as defined in section 2.13.5). PCR was run using respective primers. At the end of the cycling, the amplicons were detected in 1% agarose gel.

2.13.11. Validation of cloning by restriction digest

The plasmids were extracted by miniprep Qiagen kit as described in section 2.13.4. 20µg of plasmid was digested by *Bam*HI and *Xho*I (pGEX-6P-1) or *Hind*III and *Xho*I (pET-28c) at 37^{o}C for 2 hours. The restriction enzymes were deactivated by heating at 65^{o}C for 20 minutes and samples were loaded in 1% agarose gel containing ethidium bromide to visualize the presence of inserts.

2.13.12.Validation of cloning by DNA sequencing

The cloned plasmids (extracted by Midiprep Qiagen, kit) were sent to source biosciences to validate the presence of the cloned gene and its sequence.

2.13.12. Preservation of bacterial clones

Bacterial clones, verified for the presence of inserts by colony PCR and restriction digest, were inoculated in 20ml LB broth supplemented with 50µg/ml ampicillin (pGEX-6P-1) or 50µg/ml kanamycin (pET-28c) or 50µg/ml each of chloramphenicol and kanamycin (for *E.coli* Rosetta strain). The bacteria were incubated at 37^{o}C under shaking condition over night. Next day 1ml of the bacterial culture was dispensed in 1ml of preservation media (2% (w/v) peptone, 40% (v/v) glycerol in water) in a screw cap tube and stored at -80^{o}C.

2.14. Protein techniques

2.14.1. Protein extraction from cell lines

250µl of RIPA buffer (supplemented with protease and phosphatase inhibitor) was added to cell pellet and the pellet was resuspended by pipetting and vortexing. After giving a brief (2-3 seconds) ultrasonic pulse to shear the genomic DNA, the extract was incubated on ice for 15 minutes. The extract was centrifuged at 14,000g for 3 minutes and clear supernatant was transferred to a fresh tube. This supernatant was again centrifuged at 14,000g for 2 minutes to remove any remaining debris and clean supernatant was transferred to a fresh tube. 5µl of the supernatant was used to quantify the protein by Bradford assay and typically 50-100µg of the protein was mixed with appropriate amount of 4x protein loading dye, heated at 95^{o}C for 5 minutes, cooled at room temperature. The extracts were centrifuged to remove any insoluble material before loading onto polyacrylamide gel.

2.14.2. Protein Expression in bacteria

A loop full of glycerol stock was inoculated in 20ml of LB broth supplemented with appropriate antibiotic(s) at concentration of 50µg/ml. The inoculated broth was incubated at 37^{o}C over night in shaking condition (200rpm). This starter culture was transferred to 500ml of LB broth containing 50µg/ml concentration of ampicillin or kanamycin or each chloramphenicol+kanamycin (for p-GEX-6P-1 or pET-28C or Rosetta strain respectively) and incubated for 3 hours at 37^{o}C under shaking condition (200rpm). After three hours 500µl of IPTG (1M stock) was added in the culture and reincubated for overnight at 37^{o}C under shaking condition. Next day cells were harvested by centrifugation at 10,000g for 30 minutes and stored at -80^{o}C till further use.

2.14.3. Protein extraction from bacteria

The bacterial pellet (500ml culture) was resuspended in 20ml of lysis buffer. Complete homogenous suspension was obtained by rigorous vortexing and incubated on ice for 10 minutes. The cell suspension was sonicated thrice (while placed on ice) for 3 minutes with intervals of 1 minute. The cell suspension was centrifuged at 15,000g for 90 minutes at 4^{o}C. The clear supernatant was transferred to a fresh tube for further purification.

2.14.4. Protein purification

To purify the recombinant proteins from bacterial lysate 20ml of bacterial lysate was mixed with 500µl of Glutathione sepharose fast flow or Ni slurry (pre equilibrated with the appropriate bacterial lysis buffer) and incubated at 4^{o}C overnight. Ni slurry is strongly cationic and binds strongly with the 6x his tag attached to the recombinant protein. The mixture then placed into columns and liquid was allowed to flow through. The settled beads were washed thrice with 20ml of wash buffer-I and twice with 20ml wash buffer-II. Typically five fractions were collected by adding 500µl of elution buffer containing 10mM glutathione (GST tagged protein) or 250mM imidazole (His tagged proteins), which replaced the target protein bound to slurry. In case of gradient purification two 500µl fractions were collected with each elution buffers of different concentrations of imidazole: 50mM, 100mM, 150mM, 200mM, 250mM and 300mM. 5µl, 10µl and 40µl of the samples were separated for quantification, western blot and comassie blue gel staining respectively. All fractions were stored at -80^{o}C till further use.

2.14.5. Protein quantification

Protein quantification was done using the Bradford assay. Bradford assay is a colourimetric assay in which dye binds to the protein and changes colour (Bradford, 1976). This colour change can then directly be measured by optical density (OD) which in turn is proportional to the protein concentration of the sample. For quantification of protein, a standard curve was prepared each time using solutions of known concentration of BSA: 1, 2, 4, 8, 16µg to which, TE pH8.0 was added to a total volume of 800µl. To quantify protein in the samples, typically, 5µl of protein extract (for test) or RIPA or bacteria lysis buffer (blank) was added to 795µl of TE pH8.0. To each sample 200µl of Biorad dye was added and after immediate mixing the absorbance was noted at 595nm. A standard curve was plotted by placing absorbance values against the known concentration of BSA, which then used to quantify the unknown protein concentration of samples using their OD.

2.14.6. Protein dialysis and concentration

Selected protein fractions (typically 2nd to 5th) were poured in amicon Ultra-4 centrifugal filter units of appropriate molecular weight cut-off (for USP7; 100kDa and for EBP2; 50 and subsequently 30 kDa). After placing the samples, the units were centrifuged at 4000g at 4^{o}C till volume of the sample was reduced to 250µl. The retentate was collected in a fresh tube and stored at -80^{o}C till further use.

2.14.7. SDS-Poly Acrylamide Gel Electrophoresis (SDS-PAGE)

Polyacrylamide gel (10% resolving and 3% stacking) solution and was prepared as mentioned in table 2.10. The gel solution was poured in the assembled gel setting apparatus immediately after adding and mixing TEMED and 20% (w/v) APS solutions. A film of air barrier was overlaid on the gel solution by pouring little amount of water saturated butanol and the gel was allowed to polymerize for 30 minutes. Once resolving gel had polymerized, the butanol was decanted and washed thoroughly with dH_2O to remove any residual butanol, subsequently an appropriate comb was inserted between the gel slab. Appropriate volume of TEMED and 20% (w/v) APS was added into stacking gel solution and poured immediately. The stacking gel was also allowed to polymerize for 30 minutes. The gel was then placed in the electrophoresis apparatus and comb was removed gently. Wells formed were flushed thoroughly to remove any excess of polyacrylamide. The running buffer was poured in the gel apparatus and appropriate volumes of samples along with 10µl of protein ladder were loaded into desired wells. The samples were electrophoresed through gels at 200V for 3-4 hours.

Table 2.10. Composition of gels used in SDS-PAGE

Reagent	Stacking Gel (5%)	Resolving Gel (10%)
1.5M tris (pH8.8)	-----	12.5ml
1.0M tris (pH6.8)	3.75ml	----
0.5M EDTA	60µl	100µl
20% SDS	150µl	250µl
20% APS	300µl	500µl
TEMED	30µl	50µl
dH_20	21.96ml	27.22ml

2.14.8. Coomassie staining

To visualize the protein on acrylamide gel, gels were stained using comassie brilliant blue dye. The dye binds with protein and this complex stabilizes the anionic property of dye which in turn produces blue colour. The gels were fixed for 30 minutes in 50ml fixation solution at room temperature with gentle shaking. After fixation the solution is decanted gently and gels were stained in 50ml coomassie blue staining solution for an hour or till the whole gel turned blue. The gels were then destained using 50ml of destaining solution and later with dH_2O over night at room temperature on gentle shaking.

2.14.9. Western Blotting

For western blot, the proteins were transferred from gel to ImmobilonTM-P 0.45µm membrane (Millipore) by electrophoresis at 1500mA, 4^oC for 2 hours 20 minutes. After the transfer, the membrane was blocked using 40-50ml of blocking solution for 2 hours to over night at 4^oC. After which the membrane was probed with appropriate specific antibodies with desired dilution in membrane probing buffer and incubated for 2 hours to over night at 4^oC. If required (for instance anti His antibodies are HRP conjugated thus do not require secondary antibodies) washed the membrane with 1xPBST thrice for 10 minutes on shaking at 4^oC and probed the membrane with appropriate dilution of secondary antibodies in the membrane probing buffer for 2 hours at 4^oC. The membrane was washed thrice with 1xPBST for 10 minutes on shaking at 4^oC. The conjugated HRP was detected using ECL plus kit. The HRP conjugated enzyme oxidises the substrate (acridinium ester) in the detection kit and produces light (chemiluminescence). This resulting light is detected by autoradiography using X-ray films. The probed membrane was exposed to 8ml mixture of both reagents (1:1 ratio; provided in the ECL kit) for 45-60 seconds at room temperature. The membrane was then placed in the plastic bag and sealed. X-rays films were then placed over the membrane in the cassette for different time duration to get the appropriate impression of protein on the films.

2.14.10. Stripping of membrane

Membrane, if required to be reprobed with the same or different antibody(ies), was stripped by incubating in the stripping buffer for 1 hour at 50^{o}C in water bath. The stripped membrane was then blocked and probed as required after washing with 1xPBST three times for 10 minutes on shaker.

2.14.11. Peptide synthesis

Peptide libraries were synthesized by automatic SPOT synthesis (Kramer and Schneider-Mergener, 1998) on Whatman 50 cellulose membrane support using Fmoc (9-fluroel methoxycarbonyl) chemistry on Autospot Robot ASS222 peptide synthesizer (Invatis Bioanalytical Instruments AG, Cologene, Germany). Note, for glass slide array the peptides were synthesized separately and subsequently spotted on the sabstrum. Alanine-scanning libraries were synthesized and spotted for the peptides showing positive reaction in the glass arrays and each residue with the peptide was sequentially changed to alanine or aspartate (if alanine is a natural residue).

2.14.12. Peptide array probing and development

Peptide array was blocked with the appropriate blocking solutions (different for membrane and slide array) for four hours at room temperature on gentle shaking. Note; membrane peptide array was rinsed with absolute ethanol before blocking. After blocking, the arrays were probed with specific antibodies (appropriate dilution) or proteins (appropriate concentration) in the appropriate probing buffer (different for membrane and slide array) for overnight at 4^{o}C at gentle shaking. The array was washed thrice with 1xTBST for 10 minutes and probed with the appropriate dilution of appropriate antibodies for two hours at room temperature on gentle shaking. Array was washed thrice with 1xTBST at room temperature for 10 minutes and developed in 2ml (for slide array) or 4 ml (for membrane array) mixture of ECL detection kit solutions (after probing with appropriate secondary antibodies if required) similarly as western blot.

2.14.13. Stripping of the peptide array

The membrane array was stripped by incubating the arrays in array stripping buffer (pre heated at 70^{o}C) for 30 minutes. The array was washed twice with 1xTBST for 10 minutes at room temperature, blocked and reprobed after rinsing it with absolute ethanol as mentioned in section 2.14.12.

2.15. Cell culturing and related techniques

2.15.1. Standard cell culturing conditions

Suspension cell (B cell) lines were grown in RPMI 1640 medium, supplemented with 10% (v/v) fetal calf serum (FCS), 2% (v/v) penicillin/streptomycin (10,000 U/ml stock) solution and 2% (v/v) L-glutamine (200mM stock). All cell suspensions were incubated in vented flasks at 37^{o}C and 5% CO2.

2.15.2. Cell counting by trypan blue exclusion

Trypan blue dye is used to count the number the viable cells. Trypan blue is a polar dye that can cross the cell membrane barrier of dead or dying (necrotic and apoptotic) cells but unable to pass through the living cell membrane. This property of dye than can be used to distinguish dead or dying cells (stained) from viable (unstained) cells. The cells were diluted in appropriate ratio (e.g. 1:10) with 0.4% solution of trypan blue. 10µl of this suspension was dispensed into haemocytometer chamber and viewed under the inverted microscope. All 4 x 16 chambers were counted for live (unstained) cells. Each sample was counted in triplicate and average count was deduced from it. The average count of the cells was placed in the following equation to deduce the number of cells/ml of the suspension.

$$\text{Number of cells/ml} = (\text{Total number of cells counted}/4)\ \text{X dilution factor}\ \ \text{X}\ 10^4$$

2.15.3. Harvesting cell pellet

Cells were centrifuged at 194g for 5 minutes and supernatant was decanted. The cells were suspended with appropriate volume (0.5-1.0ml) of cell culture grade PBS (per 1x10^{6} cells for suspension cells). The suspension was centrifuged again at 2000g for 5 minutes and supernatant was decanted. Residual PBS was removed gently by pipette and cells pellet were vortexed for a short time, after which pellets were snap frozen using liquid nitrogen and stored at -80^{o}C.

2.15.4. Freezing of viable cells

A densely populated cell suspension was centrifuged at 194g for 5 minutes and supernatant was decanted. Cell pellet was resuspended in 5ml of cell culture grade PBS. Cell counting was performed and cells were centrifuged again at 194g for 5 minutes. PBS was decanted and cell pellet was resuspended in appropriate volume (based of cell density desired) of freezing medium (90% (v/v) FCS, 10% (v/v) DMSO). 1ml of this suspension was dispensed in screw cap vials and frozen slowly in an insulated box at -80^{o}C for 1-2 days before being transferred to liquid nitrogen storage.

2.15.5. Thawing of viable cells

To revive the stored cells, cells were thawed quickly at 37^{o}C and resuspended in 10 ml of prepared (supplemented) and prewarmed RPMI 1640 medium. After 1-2 hours of incubation, the cells were centrifuged at 194g for 5 minutes and the supernatant was decanted. The cell pellet was resuspended in prewarmed RPMI 1640 medium and incubated at 37^{o}C, 5% CO_2.

2.15.6. Treatment of peptides with different cell lines

Cells were centrifuged at 200g for 5 minutes and resuspended in 3 ml of RPMI medium. Cells counts were carried out by trypan blue assay and 10^6 cells were seeded in 6 well tissue culture plates. Inhibitor peptide (dissolved in cell culture grade PBS) was dispensed in all wells (to the final concentration of 25µM in 2 ml of the medium) while similar volume of only PBS was added in all control wells and the plate was left for 20 minutes at room temperature. RPMI 1640 medium was added in each well to a total volume of 2ml. After gentle rotation, plate was incubated at 37^{o}C, 5%CO_2. The 25µM of DIP (final concentration) was added in each test well for each of the 3 subsequent days. The experiment was run in triplicate.

2.15.7. DAPI staining

10µl of DAPI solution (300nM stock) was dispensed in 200µl of cell suspension and incubated at room temperature in dark for 2-3 minutes. Cells were centrifuged at 200g for 30 seconds and supernatant was removed. Cells were resuspended in 200µl of tissue culture grade PBS and centrifuged again for 30 seconds. After decanting the supernatant same procedure of washing is repeated. Finally the cells were resuspended in 50µl of tissue culture grade PBS.

2.15.8. Confocal microscopy

The cells, treated with DIP-Flu and stained with DAPI, were dispensed on glass bottom culture dishes (MatTek Cat#P35G-0-10-C) and observed under confocal microscope (Zeiss LSM 510 Meta). Flu is a green florescein molecule which excites at 488nm wave length thereby its signal is best observed at FITC channel. On confocal microscopy, the DIP-Flu was visualized using Argon/2 laser (488nm) and DAPI stained nuclei were observed with laser diode (405nm). Images were taken at different magnifications and at different planes.

2.16. Phylogenomics studies

2.16.1. Datamining

DNA and protein sequences were retrieved from NCBI (Wheeler *et al.*, 2008) and ENSEMBL (Flicek *et al.*, 2013) databases by BLASTing (blastn and blastp) against humans protein or cDNA sequences under default parameters. Additionally, the BLAST search of the homologues was further strengthened using HMMER (Finn *et al*., 2011), which employ hidden Markov model for finding homologues. Full-length EBNA1 sequences were retrieved from the UniProt database (http://www.uniprot.org/) (Magrane *et al*., 2011). To verify, each hit was reciprocally BLASTed against human genome. Moreover, in case of USPs, characteristic peptidase C19 domain of each homologue was also used to improve the search of their orthologues in selected species. In few cases where the complete cDNA sequences were not available, the predicted full coding sequence was compiled from the inferred exon structures provided on ENSEMBL. The list of sequences retrieved is provided with the organism names and accession numbers in Appendix I, IV and VI.

2.16.2. Multiple sequence alignment

Complete coding cDNA sequences (chitinase and ChiLs), full length proteins sequences (EBNA1) and peptidase C19 domain sequences (USPs) were aligned using CLUSTALX under default parameters (Thompson *et al*., 1997). Manual adjustment in the alignment (removing long unique indels) were made where necessary using Bioedit. For chitinases and ChiLs amino acid sequences representing each paralogue were also aligned using CLUSTALX after removing N-terminal signal sequences as identified by UniProt (Magrane *et al*., 2011). Where present, C-terminal tails (chitin binding domain of CHIT1 and CHIA; mucin like tail of OVGP1) were also removed. The alignment files were visualized by CLC sequence viewer.

2.16.3. Phylogenetic tree reconstruction

All phylogenetic trees were reconstructed using the maximum likelihood method from cDNA (chitinases and ChiLs) or protein sequences (EBNA1 and USPs) employing the evolutionary model (mentioned in the legends of the respective trees) selected on the basis of least Bayesian Information Criterion (BIC) value using MEGA5.20 (Tamura *et al*., 2011). Genes with incomplete/partial sequence were excluded from the analysis. Statistical support values were generated by 1000 bootstrap replicates. The consensus tree topology were developed using 50% majority rule. The substitution rate heterogeneity was incorporated by choosing Gamma distribution (based on the best fit model) to model difference in the evolutionary rates. The Nearest Neighbour Interchange heuristic method was selected to generate original trees. All phylogenetic trees of USPs family genes were rooted using bacterial (*Candidatus amoebophilus asiaticus*) protein bearing C19 domain whereas chitinase and ChiLs tree were variably rooted with homologues of *Caenorhabditis elegans* or *Branchiostoma floridae* as mentioned in the respective tree (figure) legends.

2.16.4. Evolutionary distance analyses

To estimate the evolutionary distance between one gene and a selected group of genes (usually 5 or 6), all nucleotide/protein sequences for comparison were aligned using CLUSTALX and manually adjusted where necessary. The alignment file was converted into MEGA format and the pair wise evolutionary distance was calculated using the maximum composite likelihood method (Dessimoz and Gil, 2008). The nature of data distribution was identified by Kolmogroe Smirov test. The mean of the distances was calculated and compared for the statistical significance from the unpaired student t test (normal distribution) or Wilcoxon signed rank test (skewed distribution) using Prism v4.0c.

2.16.5. Genomic synteny analyses

Genomic synteny of the selected species was investigated by locating cDNA sequence of the homologues to the genomic maps provided by NCBI and Ensembl databases. BLAST and reciprocal BLAST search of sequences of the un-annotated genes adjacent to these sites were also carried out to explore any evolutionary relationship.

2.16.6. Gene expression data

Anatomical pattern of gene expression was assessed using BioGPS database (Wu *et al.*, 2009). BioGPS is an online public database which allows comparison of gene expression based on the high density oligonucleotide microarray experiments performed uniformly for 79 human tissues.

2.16.7. Protein network analyses

Protein interaction network of USPs were assessed using STRINGv9.05 (Szklarczyk *et al.*, 2011). Unless stated otherwise, binding partners were with only high confidence threshold score of 0.7 were considered. The STRINGv9.05 database provides comprehensive coverage to both experimentally determined as well as predicted protein interactions of the query protein. The interaction map construction is primarily based on co expression, genomic context, highthrough put screening and reported empirical evidences.

2.16.8. Glycosylation site prediction

Potential glycosylation sites of the protein sequences were predicted using CBS servers (NetNglyc and NetOglyc) at a threshold value of 0.5.

2.16.9. Protein domain identification

Conserved Domain Database (CDD) of NCBI (Marchler-Baeur *et al.*, 2009) and UniProt database were exploited to identify the protein domains and catalytically active sites in USPs.

2.17. Protein structure prediction

2.17.1. Prediction of disordered regions of proteins

Protein structural propensity was predicted using FoldIndex (Prilusky *et al.*, 2005). FoldIndex is a method developed from charge-hydropathy plots (Uversky *et al.*, 2000) by changing the arrangement of the basic equation and further improvised by incorporating the sliding window. In this study, sliding window of 10 residues was used to assess the protein structural propensities.

2.17.2. Protein homology modelling

Protein homology modelling was carried out using Modeller 9v8 (Eswar *et al.*, 2006) and Swiss Model (Kiefer *et al.*, 2009). Input for the template is provided manually using the best template(s) selected by PDB Blast search. Human CHIA structure (PDBid: 3FXY) was selected as a template to model human OVGP1 and cow Chio. Similarly, murine specific ChiLs: Chil4, Chil5 and Chil6 were modelled using mouse Chil3 (PDBid: 1E9L). Other mouse homologues such as Chit1, Chia, Ovgp1, Chil1 and Chid1 were modelled using the already resolved structures of human orthologues: 1LQ0 (Chit1), 3FXY (Chia and Ovgp1), 1NWR (Chil1) and 3BXW (Chid1) respectively. Modeller constructed the protein models by satisfying the spatial restraints at both secondary structure elements and flexible loops (Sali and Blundell, 1993; Fiser *et al.*, 2003). The final model selection was based on the normalized Discrete Optimized Molecule Energy (DOPE) score .

2.17.3. Protein modelling using i-TASSER

Protein models of full length EBNA1 of human and primates lymphocryptoviruses, DIP peptide, full length human CHIT1, CHIA, OVGP1, CTBS and mouse CTBS were developed using I-TASSER (Roy *et al.*, 2010). The protein primary sequences were input to the Itasser which employs multiple threading programs using replica exchange Monte-Carlo simulation to construct structural models of the input. I-TASSER is a metaserver, which uses multiple threading programmes to get the large query coverage for templates and since 2007 to date it is the best ranked method for protein structure prediction in several benchmark studies (Battey *et al.*, 2007; Cozetto *et al.*, 2009; Zhang, 2009; predictioncenter.org/casp10/groups_analysis.cgi?type=server&tbm=on&tbm_hard=on&tbmfm=on&fm=on&submit=Filter, 2013). In the beginning, I-TASSER conducts the position specific iterated BLAST (PSI-BLAST) to identify the evolutionary relative and to generate a sequence profile (Altschul *et al.*, 1997). The sequence profile is then used to predict the secondary structure using PSI-PRED (Jones *et al.*, 1999). Subsequently the query sequence, generated sequence profile and secondary structure prediction are threaded to the PDB structure library using meta threading server LOMETS, which itself is a collection of seven threading programmes (Wu *et al*., 2007b). Each program in LOMETS finds the suitable template and ranks them on the basis of sequence and structural properties. The well aligned fragments are excised from the template while the unaligned regions of input sequence are modelled *ab-initio* (Wu *et al*., 2007a). In order to generate the full length EBNA1 model, I-TASSER automatically selected several fragment templates comprising: EBNA1 C-terminal dimer (1B3T); yeast fatty acid synthetase (2PFF), lipase (2Z8X), photosynthetic reaction centre (1C51), type A collagen (1YOF) and sineric 6-phosphoglucouronate dehydrogenase (2ZYD). Incase of CTBS, other known GH18 protein structures were selected: PDBid: 1VF8 (CHIL3), 1WB0, 1LQ0 (CHIT1), 3ALF (Class V chitinase from *Nicotinia tobaccum*), 4AY1 (CHIL1), 3FXY (CHIA) whereas C-terminal tails of CHIT1, CHIA and OVGP1 were modelled *ab-initio*. Cluster centroids were generated using replica exchange Monte-Carlo simulations (Zhang *et al*., 2002) and by averaging all the clustered structure decoys. Subsequently steric clashes were removed and global topology of the clustered centroids was improved for the second round of simulation. The external constraints were gathered from the threading alignment and PDB structures that is closest to the clustered centroids. Finally REMO (Li and Zhang, 2009) is employed to develop the full length structure of the query sequence based on the decoy generated in the second round of simulation. The models are ranked on the basis of C and TM score (Zhang, 2008; Zhang and Skolnick, 2004) to assess the similarity between the target protein model and known template structures.

2.17.4. Protein modelling using Molecular Operating Environment (MOE)

Models were also generated in MOE using the template 1B3T (with and without DNA) for homology modelling and *ad hoc* outgap modelling to similar fragments from PDB for the remainder of the sequence. An initial proposed partial geometry was copied from the template chains in the solved structure of 1B3T by using all coordinates where residue identity was conserved. Otherwise, only backbone coordinates were used. Based on this initial partial geometry, Boltzmann-weighted randomized modelling (Levitt, 1992) was employed with segment searching in PDB for regions that could not be mapped onto the initial partial geometry (Fechteler *et al*., 1995). Twenty-five models were constructed. On completion of segment addition, each model was energetically minimized in the AMBER-99 force field (Wang *et al*., 2000). The highest-scoring intermediate model was then determined by the generalized Born/volume integral (GB/VI) methodology (Labute, 2008). All modelling using MOE was conducted by Dr. Derek Gatherer.

2.17.5. Structural assessments of models

The models were assessed for their structural plausibility by generating Ramachandran plots and detecting bad angles using Molprobity (Chen *et al.*, 2010). Additionally, normalized Q mean score were estimated for each model using Q mean server to assess the quality of models (Benkert *et al.*, 2009). Normalized Q mean score is a method to derive the global and local error estimates independent to the protein length. The structural and spatial variability in the Cα back bone and orientation of ligand binding residues were assessed in terms of Root Mean Square Deviation (RMSD) by superimposing one structure over the other using Swiss-PdbViewer v4.0.2 (Johanson *et al.*, 2012). All structures were visualized and electrostatic surface were generated using DS visualizer v3.5.

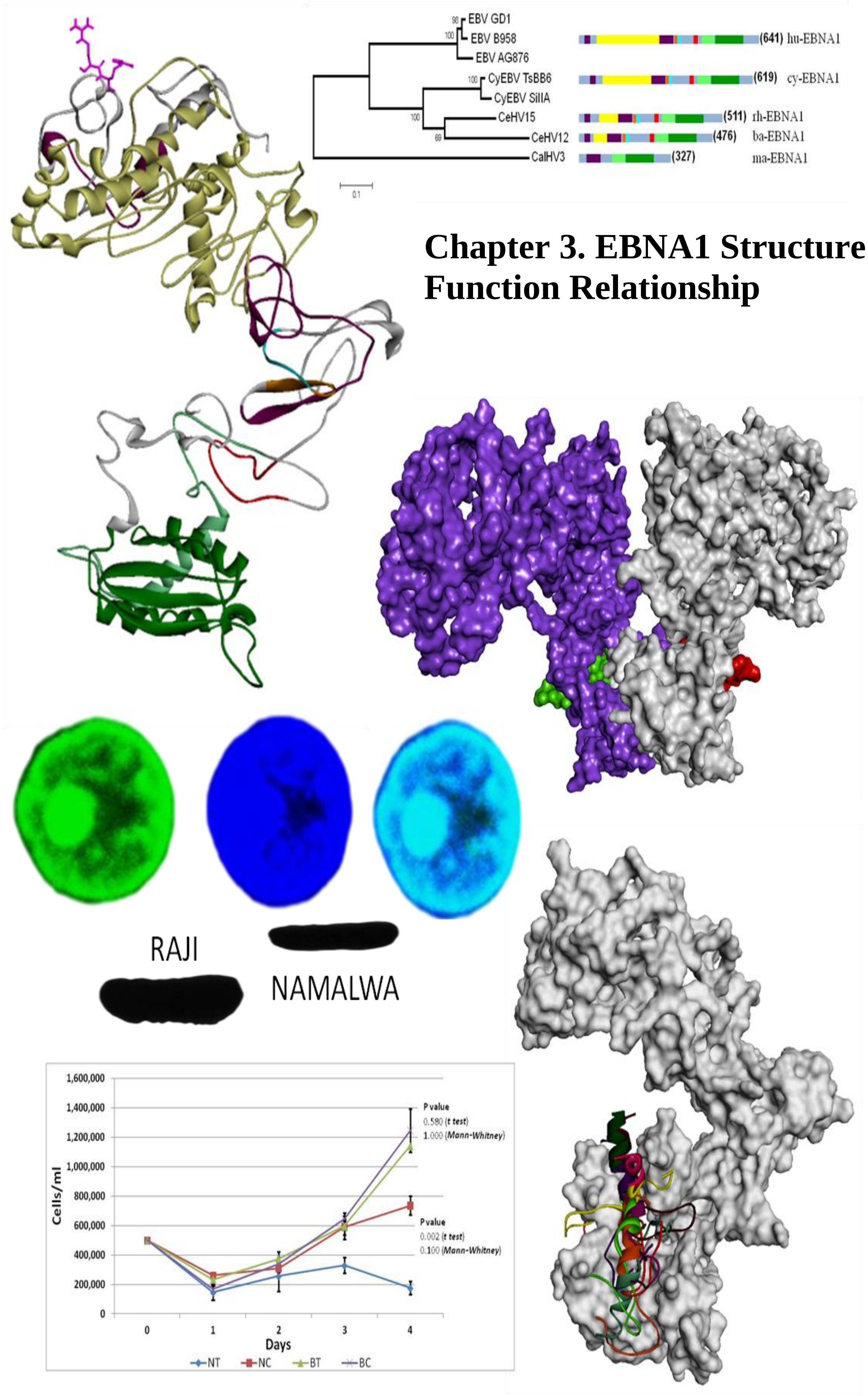
2.17.6. Cavity analyses

Protein cavity (ligand binding groove) volumes were calculated using POCASA 1.0 (Yu *et al.*, 2010). Briefly, the program scans the protein in a 3D grid with 2.0Å probes for the cavities. When the search is completed the cavities are automatically displayed on the structure and ranked on the basis of volume (accumulation of the spherical probes).

2.17.7. Molecular Docking

Molecular docking was performed using SymmDock (Schneidman-Duhovny *et al.*, 2005b) and ClusPro2.0 (Comeau *et al.*, 2004) adopting both blind and directed (by providing interaction sites information) approaches. Symmdock is exploited to predict the EBNA1 dimer conformation. The program exploits the local features of protein (input) to produce symmetric cyclic transformation at a given order n (for dimer; n=2). This symmetric cyclic transformation is later being exploited for clustering and finally to predict the dimer conformations of a given monomer (Schneidman-Duhovny *et al.*, 2005b). Conformations are ranked on the basis of geometric score, desolvation energy (Zhang *et al.*, 1997) and the interface area size. According to bench mark studies the near native conformation are usually found among the top 20 ranked predictions (Chen *et al.*, 2003; Inbar *et al.*, 2005). In this study we analysed first 20 ranked conformations of EBNA1 dimer and decision was made on the majority rule. ClusPro2.0 is currently the highest ranking algorithm for its reliability for the structural prediction of protein protein interaction (Kozakov *et al.*, 2010) and it was used to predict the structural conformation of EBNA1-DIP interactions. ClusPro uses Fourier transform correlation technique based docking platforms; DOT and Z dock to predict the protein-protein interaction conformation. Scoring of the predicted conformation (depending on the platform) is based on pairwise shape complementarity (PSC), desolvation energy values and electrostatic values. The programme filters top 2000 docking simulations on the basis of pairwise binding site RMSD criterion to yield the final structure which is the further refined by applying CHARMM forcefield.

Chapter 3. EBNA1 Structure Function Relationship



3.Results: EBNA1 Structure-Function Relationship

3.1. Introduction

Our understanding of EBNA1 structure is fragmented, as to date only the C-terminal portion of the protein, involved in DNA binding and dimerisation, has been structurally resolved (Bochkarev et al., 1995). The remaining one third of the protein which include glycine alanine repeat region (GAr domain) and glycine arginine rich regions (GR1 and GR2) and other host protein interacting regions have not been structurally resolved (Figure 1.4). Given many of the EBNA1 functions such as transactivation, genome maintenance and conferring resistance to apoptosis are regulated with these regions; understanding the structural biology of these regions is of considerable significance. This chapter describes the structure-function relationship of EBNA1 encoded by EBV and other primate LCVs. In order to explore structural and/or functional divergence in EBNA1 of different LCVs, molecular models of EBNA1 protein for all LCVs were constructed and compared in the monomeric and dimeric conformations. Based on the observations, a peptide (DIP) was proposed that could interfere with the EBNA1 dimer stability. In the present study the preliminary evaluation of the efficiency of DIP was evaluated.

3.2. Phylogenetic analysis of EBNA1

With the exception of the GAr region, EBNA1 shows limited sequence identity to any other proteins in the databases. However, several homologues of EBNA exist in different herpes viruses especially LCVs that infect primates. Several proteins were identified by textmining in rice (*Oryzia sativa*) and bacterium (*Erwinia chrysanthemi*) denoted as EBNA1, EBNA1-like or EBNA1-nuclear protein but with little or no sequence identity to herpesvirus EBNA1. Moreover, reciprocal BLAST did not show any relationship of these sequences with herpes viruses EBNA1.

Since the GAr region is composed of a relatively simple stretch of residues, to improve the accuracy of the BLAST search, this region was deleted for the BLAST search. In total 8 complete homologues were retrieved: 3 homologues from different strains of EBV and single homologues each from LCV infecting cynomolgus monkey (*Macaca fasicularis*), rhesus macaque (*Macaca mulatta*), baboon (genus Papio) and marmoset (*Callithrix jacchus*). These sequences are denoted as hu-EBNA1 (from EBV/HHV4), cy-EBNA1 (from CyEBV), rh-EBNA1 (from CeHV15), ba-EBNA1 (from CeHV12) and ma-EBNA1 (from CyEBV), rh-EBNA1 (from CeHV15), ba-EBNA1 (from CeHV12) and ma-EBNA1 (from CalHV3). The reconstructed phylogenetic tree using these sequences revealed the separation of the single New World primate EBNA1 sequence (ma-EBNA1) from the Old

World primate virus sequences (Figure 3.1). Expectedly, human sequences were clustered together and other Old World primates sequences formed a distinct subclade. The overall tree topology resembles the extensive phylogenetic reconstruction based on the LCV glycoprotein B genes (Duellman *et al*., 2009).

3.3. Multiple sequence alignment of EBNA1

To compare the primary protein structure of EBNA1 homologues, primate LCVs EBNA1 sequences were aligned, using EBNA1 of EBV B95-8 strain as the reference protein (Figure 3.2). The hu-EBNA1 sequences are the longest homologues of EBNA1 and show 88% to 97% identity to each other. Identity between the Old World monkey LCV EBNA1 is 35% to 46%, while ma-EBNA1, the shortest homologue of EBNA1, shows the limited identity with the other sequences (Table 3.1). The GAr of hu-EBNA1 spans over one third (90-324) of the total protein length and is predominantly composed of Gly and Ala residues (with the exception of Glu at position 273 and 274 in GD1 strain of human EBV). In comparison the GAr region of the Old World monkey LCVs is shorter and intervened by other residue (predominantly Ser and Val) while it is completely absent in EBNA1 of CalHV3 (Figure 3.2).

The GAr region (where present) is flanked by Gly and Arg repeat regions (GR1 and GR2), which are present in EBNA1 homologues of all human and Old World monkey viruses, suggesting a conservation of function and reflecting their involvement in the viral genome replication, maintenance and transactivation. By contrast, ma-EBNA1 has only one GR region which shares higher sequence similarity with GR2 than GR1. Given that the GAr sequence is absent, a single GR region of ma-EBNA1 might reflect a domain that has not been cleaved by the GAr (Figure 3.1). A stretch of 10 amino acids (KRPSCIGCKG), just C-terminal to GR1 in hu-EBNA1 is strongly conserved in Old World primate LCVs EBNA1 suggesting conserved biological role, however, the region is absent from ma-EBNA1. Two Cys residues in the N-terminal region of ma-EBNA1 (residues 38 and 43) aligned to Cys79 and Cys82 of hu-EBNA1 (Figure 3.2). These cysteines may perform a similar function to Cys79 and Cys82 in hu-EBNA1, within the highly conserved stretch in Old World primate virus EBNA1's.

The interaction sites for USP7 and CK2 on hu-EBNA1 are conserved in Old World monkey virus EBNA1 homologues. In particular, residues involved in intermolecular hydrogen bonding between EBNA1 and USP7 (EBNA1 Pro442, Glu444, Gly445 and Ser 447) (Saridakis *et al*., 2005) are present at the corresponding position in EBNA1 of Old World monkey viruses. However, ma-EBNA1 did not show a similar sequence at the

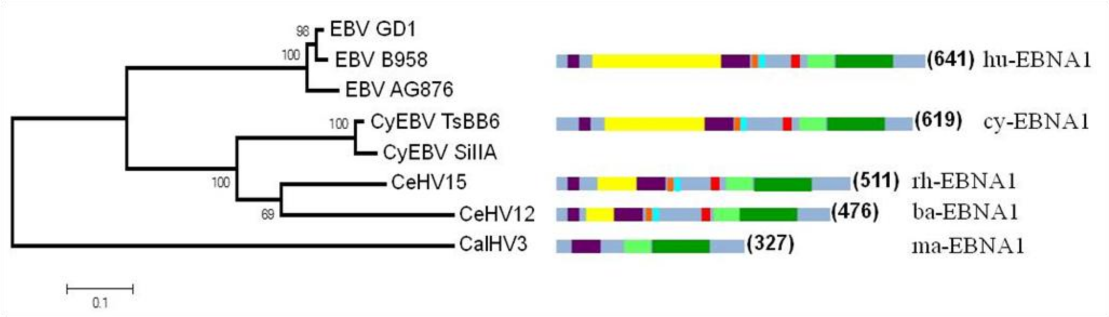
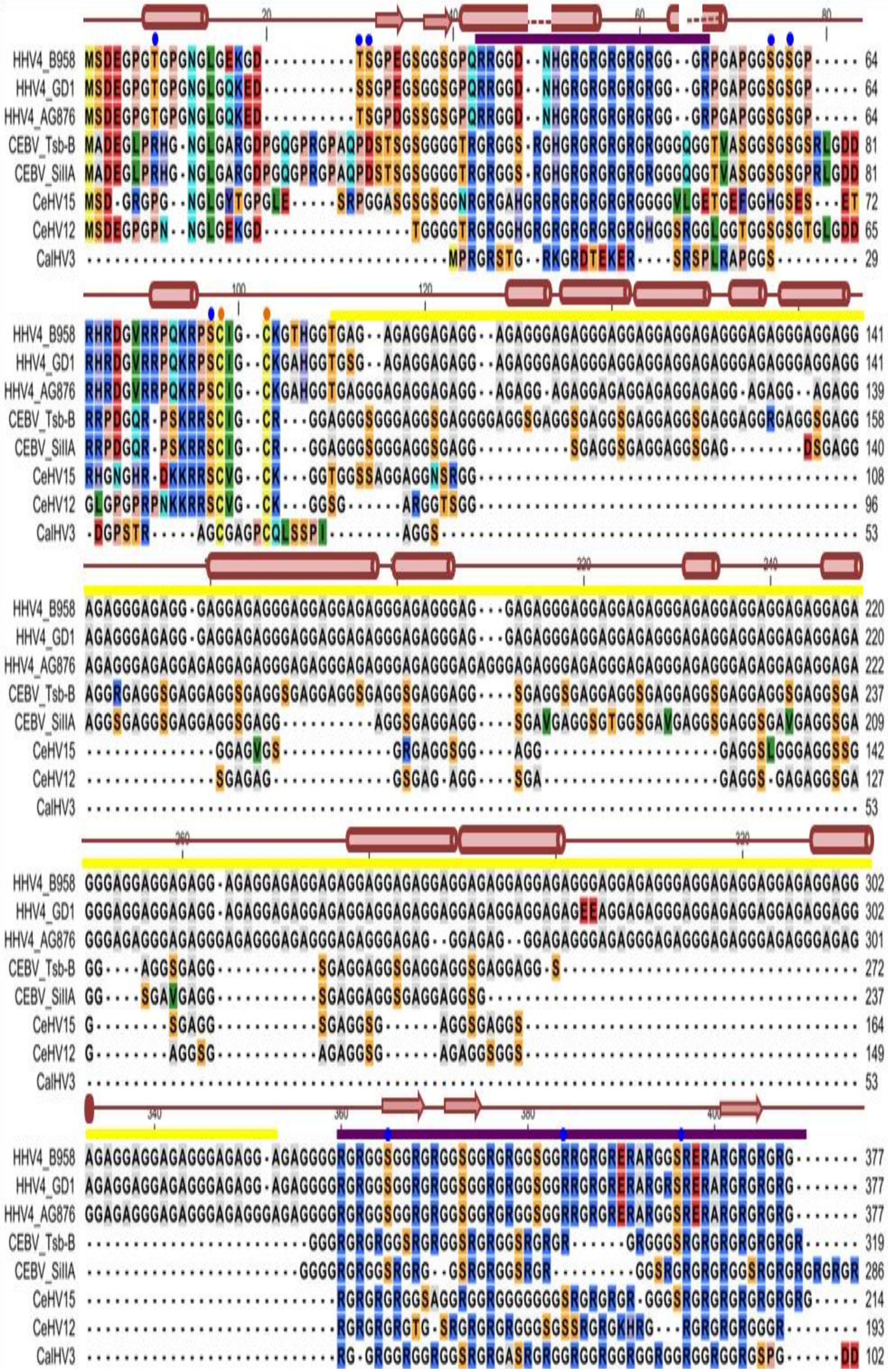


Figure 3.1. Phylogenetic tree of EBNA1 homologues. The evolutionary history of primate LCV EBNA1 homologues was reconstructed using protein sequences and the maximum likelihood method with Whelan and Goldman replacement model (Whelan and Goldman, 2001). The consensus tree was developed with 1000 bootstrap replicates. One clade includes three EBV (HHV4) homologues of EBNA1 (huEBNA1). Non-human Old World monkey LCVs EBNA1 form a separate clade (herpes viruses of cynomolgus monkey (CyEBV), rhesus macaque (CeHV15) and baboon (CeHV12), while the marmoset CalHV3 EBNA1 homologue outgroups both clades. The domain organization of homologues is schematically represented on the right with protein length indicated in brackets. The predicted protein domains or sites are colour coded: purple: Gly, Arg repeat region (GR1 and GR2); yellow: GAr; orange: NLS; cyan: CK binding site; red: USP7 binding site; green: DNA binding domain (core in dark green).

	HHV4 B958	**HHV4 GD1**	**HHV4 AG876**	**CEBV TsbB6**	**CEBV SiIIA**	**CeHV15**	**CeHV12**	**CalHV3**
B958		97	88	46	44	38	36	22
GD1			87	46	43	38	36	22
AG876				44	42	37	35	21
TsbB6					90	54	47	21
SiIIA						56	50	23
CeHV15							59	25
CeHV12								26
CalHV3								

Table 3.1. Percentage identity of EBNA1 homologues. Sequence identities between different EBNA1 homologues are shown in percentages, based on the multiple sequence alignment.


20
40
60
80
HHV4_B958 MSDEGPGTGPGNGLGEKGD..........TSGPEGSGGSGPQRRGGD..NHGRGRGRGRGRGG..GRPGAPGGSGSGP..... 64
HHV4_GD1 MSDEGPGTGPGNGLGQKED..........SSGPEGSGGSGPQRRGGD..NHGRGRGRGRGRGG..GRPGAPGGSGSGP..... 64
HHV4_AG876 MSDEGPGTGPGNGLGQKED..........TSGPDGSSGSGPQRRGGD..NHGRGRGRGRGRGG..GRPGAPGGSGSGP..... 64
CEBV_Tsb-B MADEGLPRHG.NGLGARGDPGQGPRGPAQPDSTSGSGGGGTRGRGGS.RGHGRGRGRGRGRGGGQGGTVASGGSGSGSRLGDD 81
CEBV_SiIIA MADEGLPRHG.NGLGARGDPGQGPRGPAQPDSTSGSGGGGTRGRGGS.RGHGRGRGRGRGRGGGQGGTVASGGSGSGPRLGDD 81
CeHV15 MSD.GRGPG..NGLGYTGPGLE.....SRPGGASGSGSGGNRGRGAHGRGRGRGRGRGRGGGGVLGETGEFGGHGSES...ET 72
CeHV12 MSDEGPGPN..NGLGEKGD................TGGGGTRGRGGHGRGRGRGRGRGRGHGGSRGGLGGTGGSGSGTGLGDD 65
CalHV3 ....MPRGRSTG..RKGRDTEKER....SRSPLRAPGGS......... 29
100
120
HHV4_B958 RHRDGVRRPQKRPSCIG..CKGTHGGTGAG..AGAGGAGAGG..AGAGGGAGAGGGAGGAGGAGGAGAGGGAGAGGGAGGAGG 141
HHV4_GD1 RHRDGVRRPQKRPSCIG..CKGAHGGTGSG..AGAGGAGAGG..AGAGGGAGAGGGAGGAGGAGGAGAGGGAGAGGGAGGAGG 141
HHV4_AG876 RHRDGVRRPQKRPSCIG..CKGAHGGTGAGGGAGAGGAGAGG..AGAGG.AGAGGAGAGGAGAGGAGAGG.AGAGG..AGAGG 139
CEBV_Tsb-B RRPDGQR.PSKRRSCIG..CR...GGAGGGSGGGAGGSGAGGGGAGGSGAGGSGAGGSGAGGAGGSGAGGAGGRGAGGSGAGG 158
CEBV_SiIIA RRPDGQR.PSKRRSCIG..CR...GGAGGGSGGGAGGSGAGG..........SGAGGSGAGGAGGSGAG........DSGAGG 140
CeHV15 RHGNGHR.DKKRRSCVG..CK...GGTGGSSAGGAGGNSRGG... 108
CeHV12 GLGPGPRPNKKRRSCVG..CK...GGSG......ARGGTSGG... 96
CalHV3 .DGPSTR.....AGCGAGPCQLSSPI........AGGS... 53
220
240
HHV4_B958 AGAGGGAGAGG.GAGGAGAGGGAGGAGGAGAGGGAGAGGGAG...GAGAGGGAGGAGGAGAGGGAGAGGAGGAGGAGAGGAGA 220
HHV4_GD1 AGAGGGAGAGG.GAGGAGAGGGAGGAGGAGAGGGAGAGGGAG...GAGAGGGAGGAGGAGAGGGAGAGGAGGAGGAGAGGAGA 220
HHV4_AG876 AGAGGGAGAGGAGAGGAGAGGGAGAGGGAGAGGGAGAGGGAGAGGGAGAGGGAGAGGGAGAGGGAGAGGGAGAGGAGAGGAGA 222
CEBV_Tsb-B AGGRGAGGSGAGGAGGSGAGGSGAGGAGGSGAGGSGAGGAGG....SGAGGSGAGGAGGSGAGGAGGSGAGGAGGSGAGGSGA 237
CEBV_SiIIA AGGSGAGGSGAGGAGGSGAGG..........AGGSGAGGAGG....SGAVGAGGSGTGGSGAVGAGGSGAGGSGAVGAGGSGA 209
CeHV15 ..............GGAGVGS............GRGAGGSGG....AGG...................GAGGSLGGGAGGSSG 142
CeHV12 ..............SGAGAG.............GSGAG.AGG....SGA...................GAGGS.GAGAGGSGA 127
CalHV3 53
260
320
HHV4_B958 GGGAGGAGGAGAGG.AGAGGAGAGGAGAGGAGGAGAGGAGGAGAGGAGGAGAGGGAGGAGAGGGAGGAGAGGAGGAGAGGAGG 302
HHV4_GD1 GGGAGGAGGAGAGG.AGAGGAGAGGAGAGGAGGAGAGGAGGAGAGGAGGAGAGEEAGGAGAGGGAGGAGAGGAGGAGAGGAGG 302
HHV4_AG876 GGGAGAGGGAGAGGGAGAGGGAGAGGGAGAGGGAGAG..GGAGAG..GGAGAGGGAGAGGGAGAGGGAGAGGGAGAGGGAGAG 301
CEBV_Tsb-B GG....AGGSGAGG...........SGAGGAGGSGAGGAGGSGAGGAGG.S................................ 272
CEBV_SiIIA GG....SGAVGAGG...........SGAGGAGGSGAGGAGGSG.. 237
CeHV15 G........SGAGG...........SGAGGSG......AGGSGAGGS.................................... 164
CeHV12 G........AGGSG...........AGAGGSG......AGAGGSGGS.................................... 149
CalHV3 53
340
360
380
400
HHV4_B958 AGAGGAGGAGAGGGAGAGG.AGAGGGGRGRGGSGGRGRGGSGGRGRGGSGGRRGRGRERARGGSRERARGRGRGRG....... 377
HHV4_GD1 AGAGGAGGAGAGGGAGAGG.AGAGGGGRGRGGSGGRGRGGSGGRGRGGSGGRRGRGRERARGRSRERARGRGRGRG....... 377
HHV4_AG876 GGAGAGGGAGAGGGAGAGGGAGAGGGGRGRGGSGGRGRGGSGGRGRGGSGGRRGRGRERARGGSRERARGRGRGRG....... 377
CEBV_Tsb-B ........................GGGRGRGRGGSRGRGGSRGRGGSRGRGR......GRGGGSRGRGRGRGRGRGR...... 319
CEBV_SiIIA .......................GGGGRGRGGSRGRG..GSRGRGGSRGR.........GGSRGRGRGRGGSRGRGRGRGRGR 286
CeHV15 ...........................RGRGRGRGGSAGGRGGRGGGGGGGSRGRGRGR.GGGSRGRGRGRGRGRGRG..... 214
CeHV12 ...........................RGRGRGRGTG.SRGRGRGRGGGSGSSRGRGKHRG...RGRGRGRGGGR........ 193
CalHV3 ...........................RG.GRGGRGGRGGSRGRGASRGRGGRGGRGGRGGRGGRGGRGSPG......DD 102

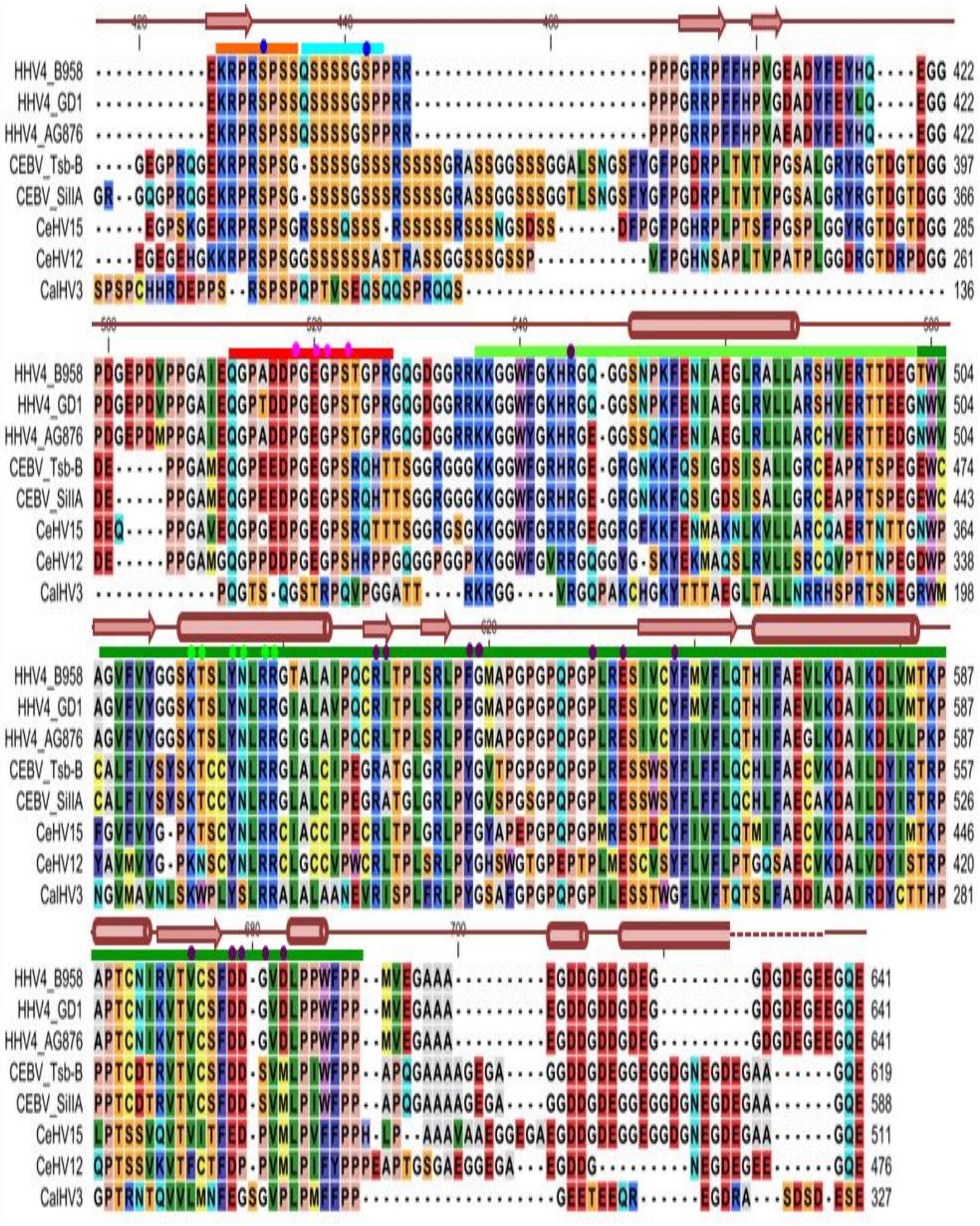


Figure 3.2. Multiple sequence alignment of EBNA1 homologues. The protein sequence of EBNA1 homologues is shown with RasMol colour coding of the residues. The secondary structural elements (based on the composite model of EBNA1 of EBV-B95-8) are shown as cylinders (α helices) and arrows (β sheets). Important protein domains or interaction sites are represented by coloured horizontal bars: purple: GR1 and GR2, yellow: GAr, orange: NLS, cyan: CK2 binding site, red: USP7 binding site, green: DNA binding domain (core domain in dark green). Coloured dots above the sequences indicated critical residues of structural and/or functional importance: blue: predicted phosphorylation sites; pink: critical residues involved in USP7 binding; purple: dimerisation; green: DNA binding; orange: conserved Cys residues at N-terminal.

corresponding region or at any other position in the alignment. Compared to hu-EBNA1, an additional stretch of Ser at the CK2 binding site is present in the Old World monkey LCV EBNA1 homologues. At the aligned region in ma-EBNA1, several Ser/Pro residues are present, it is not clear if these could constitute a CK2 binding site. This implies that either both (USP7 and CK2) binding sites have been lost during the evolutionary course of New World Monkey virus or gained during Old World monkeys LCVs evolution. A sequentially conserved stretch of approximately 30 residues is present between CK2 and USP7 binding sites in EBNA1 homologues of Old World monkey viruses but absent in the hu-EBNA1 sequences.

A single ubiquitination site on hu-EBNA1 (Lys477) was predicted, which is conserved in all the homologues. Six of the ten proposed phosphorylation sites of hu-EBNA1 (Duellman *et al.*, 2009) are also conserved in primate virus homologues (Ser60, Ser62, Ser78, Ser365, Ser383, Ser393) (Figure 3.2). The nuclear localization signal sequence (NLS: 379-385, KRPRSPS) is fully conserved in all Old World monkey LCV EBNA1 homologues, while in ma-EBNA1 only the last four residues are present at the corresponding position in the alignment. A consensus NLS sequence (K, K/R, x, K/R) is also not present in ma-EBNA1, however RKxRxxxK towards the N-terminus or RKRxxxxR at the N-terminal end of the DNA binding region might function as a NLS. In ba-EBNA1 and rh-EBNA1, a conserved KKRRS within the LR1 homology region could serve as second NLS.

The DNA binding and dimerisation domain (459-607) is the most sequentially conserved domain between the EBNA1 homologues (Figure 3.2). With few exceptions, residues shown for their involvement in DNA binding (Lys514, Thr515, Tyr518, Asn519, Arg521 and Arg522) and dimerisation (Arg469, Tyr510, Arg532, Leu533, Phe541, Gly542, Pro553, Glu556, Tyr561, Val597, Ser599, Asp601, and Asp605) (Bochkarev *et al.*, 1995) are highly conserved. Such retention of critical residues in this region suggests the structural and functional conservation between EBNA1 homologues of these primate viruses.

3.4. EBNA1 monomeric model

To understand the structural aspects of the EBNA1 biology, a full length model of the EBNA1 protein was constructed. First, the propensity of the protein molecules to adopt secondary or tertiary structure was explored by using FoldIndex web server. The FoldIndex analyses of EBNA1 homologues predicts that the N-terminal and central regions of the protein are unstructured or unfolded (Figure 3.3A). Conversely, the GAr domain is predicted to be folded and DNA binding and dimerisation domain yielded positive signals of structured conformation.

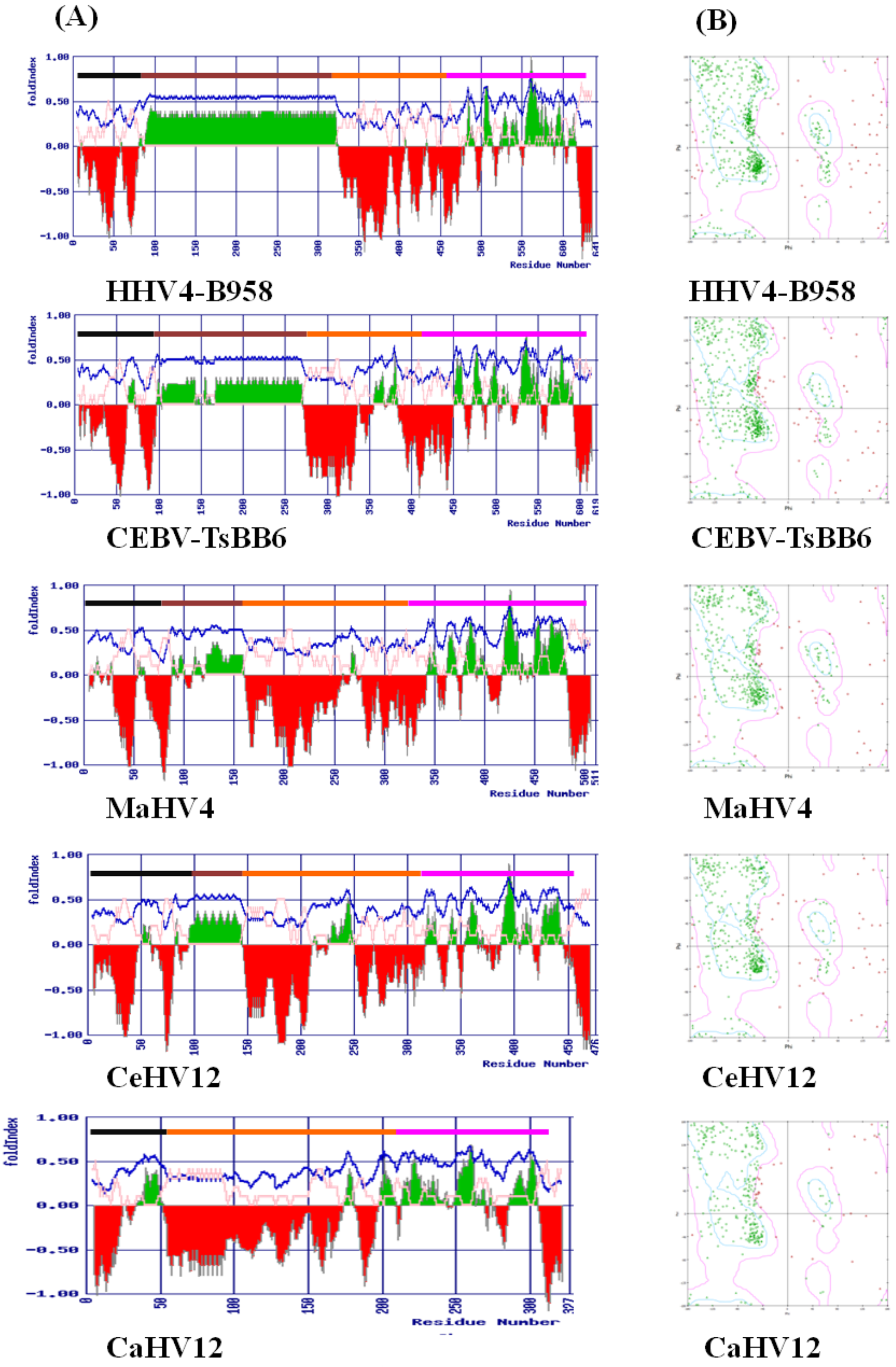


Figure 3.3. Structural propensity of EBNA1 homologues. EBNA1 protein sequences from the primate LCVs (as indicated) were used to predict the structural propensity by FoldIndex. The distribution of hydrophobic and charged residues across the protein length is shown by blue and pink lines (respectively). Potential disordered and ordered regions of protein are represented by negative (red) or positive (green) values (respectively). A simplified domain distribution is indicated as coloured bars above each plot: black: N-terminal; brown: GAr; gold: GR2 and protein binding sites; pink: DNA binding and dimerisation. Numbers on X-axis correspond the length of the protein. **(B) Ramachandran plots of EBNA1 modelled structures.** The human and other primate LCV EBNA1 protein structure models (as indicated) were evaluated for dihedral angle distribution using Ramachandran plots. Residues in allowed and disallowed regions are represented by green and pink spots (respectively). Generously and strictly allowed regions are depicted by fuchsia and cyan contour lines (respectively).

To construct the full length structural model of EBNA1 *in silico,* the hu-EBNA1 B95-8 protein was screened for any structural similarities with deposited protein structures in the RCSB database using PDB BLAST. Apart from the resolved C-terminal portion of the EBNA1, none of the available structures thus identified alone or in combination, covered the query sequence to any extent. Therefore a combined approach was adopted to exploit the advantages of three programmes (I-TASSER, MOE and Modeller) of protein modelling. Over the last five years I-TASSER has been ranked first in the CASP (critical assessment of protein structure prediction) contest for protein structure prediction (Kryshtafovch *et al*., 2009). I-TASSER employed homology modelling to model the C-terminus of the query sequence using the resolved C-terminal domain of EBNA1 (1B3T) as a template. For structurally unresolved regions, I-TASSER searches the protein structure databases (primarily RCSB) for small regions of sequential and structural similarity and uses these fragments from multiple templates to model the aligned regions. *Ab-initio* modelling procedures are employed to model unaligned regions. Finally, I-TASSER assembles all fragments of predicted structure to generate thousands of models by iterative threading which are subsequently evaluated for the best fit. Homology modelling can also be conducted using MOE, which allows the placement of spatial constraints for heteroatoms or molecules (such as DNA). In order to exploit the advantage offered by both programmes, primary models were developed separately in I-TASSER and MOE using the structurally resolved region of EBNA1 (1B3T) as a template. The best models selected from each programme were then used as a template to construct composite models. The best composite model was then selected on the basis of dihedral bond angle ratio (89.5% in allowed region and additional 5.9% in the generously allowed region) (Figure 3.3B) and lowest free energy. Owing to the relative proportions of Gly and Pro compared to other residues, these values lie within acceptable limits of protein structure. As expected, the C-terminal region (DNA binding and dimerisation domain) of the *in silico* model is nearly identical to the 1B3T crystal structure (RMSD deviation of 1.29Å) (Figure 3.4).

The EBNA1 model predicts a helix (31-43) with GR1 (part of EBP2 and RNA binding region) (Figure 3.4). A conserved residue (Arg71 and to less extent Arg72) within LR1 transactivation domain protrudes from the structure. The strong conservation and spatial position of the residue in the predicted structure suggests its biological importance. Consistent with the FoldIndex prediction, the GAr region forms multiple helices of variable length and the remainder of N-terminal region appears unstructured. Similarly, the central region (known for protein-protein interactions: EBP2 (residues: 325-476), CK2 (residues: 399-395), USP7 (residues: 436-450)) is largely unstructured except for the presence of short parallel β sheets in GR2 (residues: 334-338, 348-351 and 370-374).

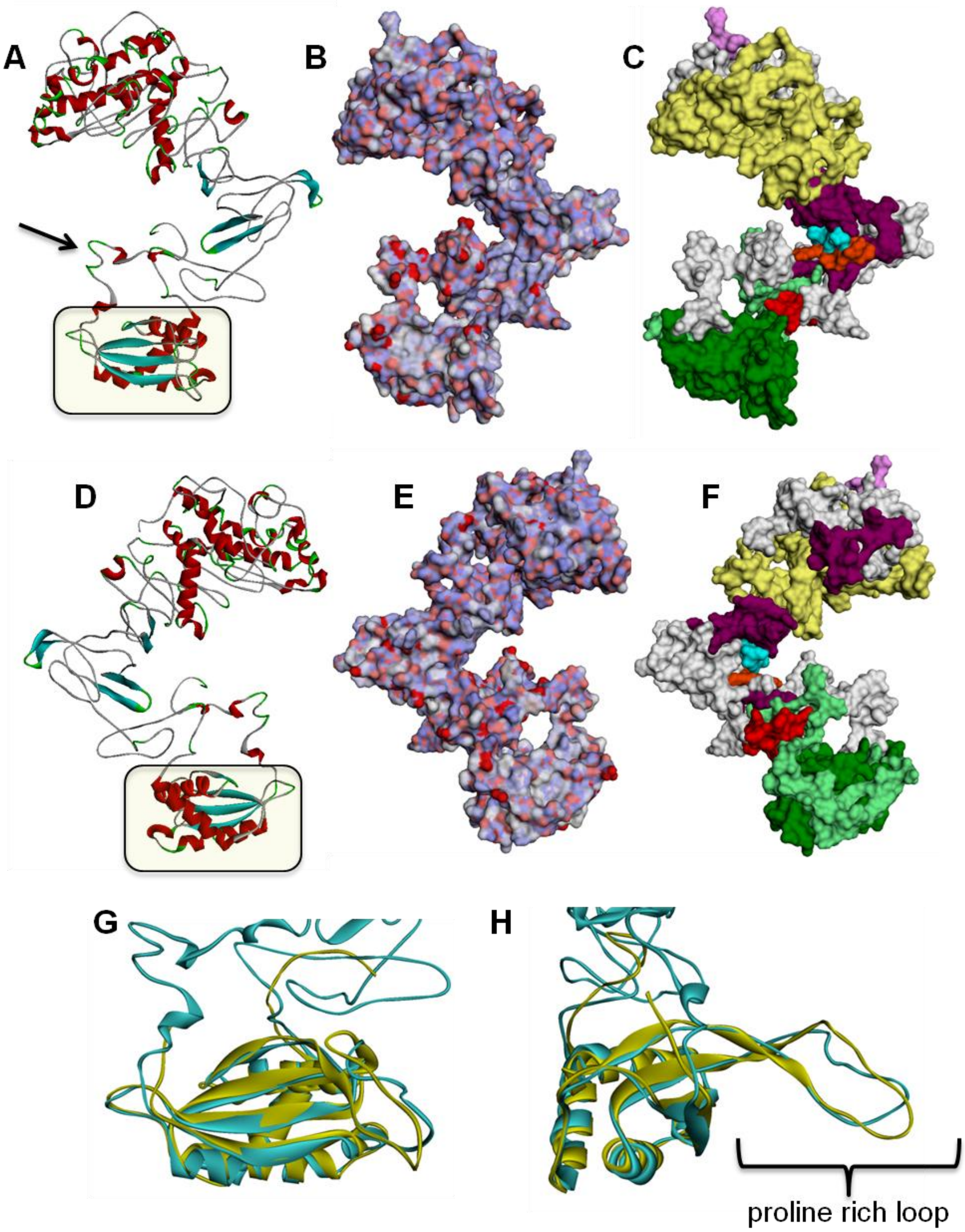


Figure 3.4. EBV EBNA1 composite model structure. A composite model of EBNA1 of EBV/B95-8, constructed using I-TASSER, MOE and Modeller is represented in (A-C), with 180^o rotation in horizontal plane (D-F). The model is shown in ribbon format (A, D) (with the structurally resolved portion boxed), electrostatic surface topology (B, E) and surface highlighting structurally and functionally important regions (C, F): yellow: GAr; purple: GR1 and GR2; pink surface: Arg71 and Arg72; cyan: CK2 interaction region; orange: NLS; red: USP7 binding site; light green and dark green: flanking region and core DNA binding and dimerisation domain respectively. Note the prehensile C-terminal tail curls back towards DNA binding region to form a ring (arrow in A). Comparison of the C-terminal region of the EBNA1 composite model and the resolved structure is shown in (G) & (H) with a horizontal rotation of 90^o. The C-terminal region of the composite EBNA1 model (cyan) is shown superimposed over a monomer retrieved from the structurally resolved template 1B3T (yellow) (RMSD value is 1.29Å). Note: the protruding proline rich loop is highlighted in (H).

The NLS region (residues: 381-385) is also modelled as disordered. Although the protein interacting region exhibits the characteristics of an intrinsically disordered region, it is possible that the region upon interacting with different partner proteins dynamically adopts alternate structural conformations. A proline rich region (residues 537-559) is highly conserved in all primate LCV EBNA1 homologues and forms a loop which protrudes in a different plane (Figure 3.4). Consistent with the FoldIndex prediction, the C-terminal tail (residues: 608-641) is largely unstructured in the predicted model except for two small helices. In the predicted model, the C-terminal tail curls back towards the DNA binding and dimerisation domain, making a ring-like conformation with a central hole (Figure 3.4).

Comparison of primate LCV EBNA1 models with the hu-EBNA1 model shows clear structural similarity at the C-terminal DNA binding and dimerisation domain (Figure 3.5). Similarly, the C-terminal tail also curls back to the DNA binding region to form a loop with hole in the middle. However, the remaining portions of the modelled proteins show several differences with hu-EBNA1. The GR1 and GR2 domain become closer due to the reduction of GAr domain length in EBNA1 of non human primate LCVs. The GAr domain is absent and a longer α helix is predicted at the N-terminus of the ma-EBNA1 model. Both USP7 and EBP2 binding sites are relatively more unstructured (especially the latter) than in hu-EBNA1. By contrast, the CK2 binding region and the intervening sequence between CK2 and USP7 binding site, form β sheets, this is consistent with the FoldIndex prediction. The prominent Arg71 and to some extent Arg72 are not predicted to protrude in the primates LCV EBNA1 molecules except in the ba-EBNA1 model where only Arg71 shows an exteriorly located side chain.

3.5. EBNA1 homodimer model

Two approaches were used to construct *in silico* homodimer models of full length EBNA1: dimer homology modelling using MOE and dimer generation in SymmDock. In the former approach, the resolved EBNA1 C-terminal domain dimer (1B3T) was used as a template to construct a full length EBNA1 dimer with DNA (Figure 3.6). Spatial restraints (manually placed) allows incorporation of atomic coordinates of the DNA molecule (as available in 1B3T) into the model and the resulting dimer model is highly similar to 1B3T in the C-terminal region (RMSD=0.35Å). In this predicted dimer, the string of residues that connects C-terminal domain with the N-terminal region of the protein neatly occupies the major groove of the bound DNA (Figure 3.6). However, the N-terminal and C-terminal ends of the model predicted solely by MOE show several differences to the composite model (Figure 3.7). Moreover, the MOE monomers in the MOE predicted dimer are not symmetrical and show differences in the distribution of secondary structural elements. This

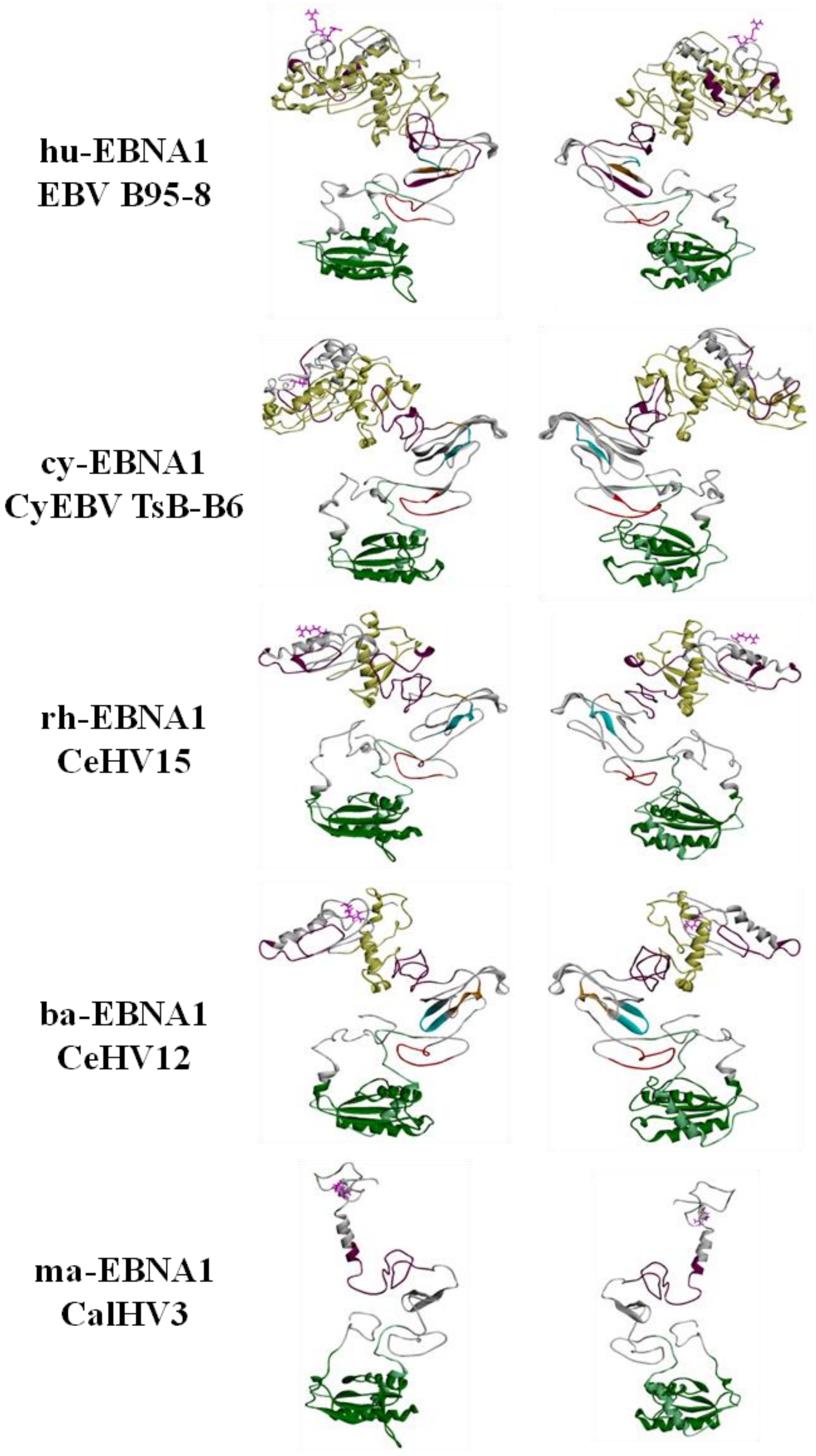


Figure 3.5. Primates LCV EBNA1 structures. Ribbon cartoons of different primate LCV EBNA1 monomers (as labelled) are shown in 180^{o} rotation in horizontal plane. Different regions are coloured differently to highlight structurally and/or functionally important regions as in Figure 3.4.

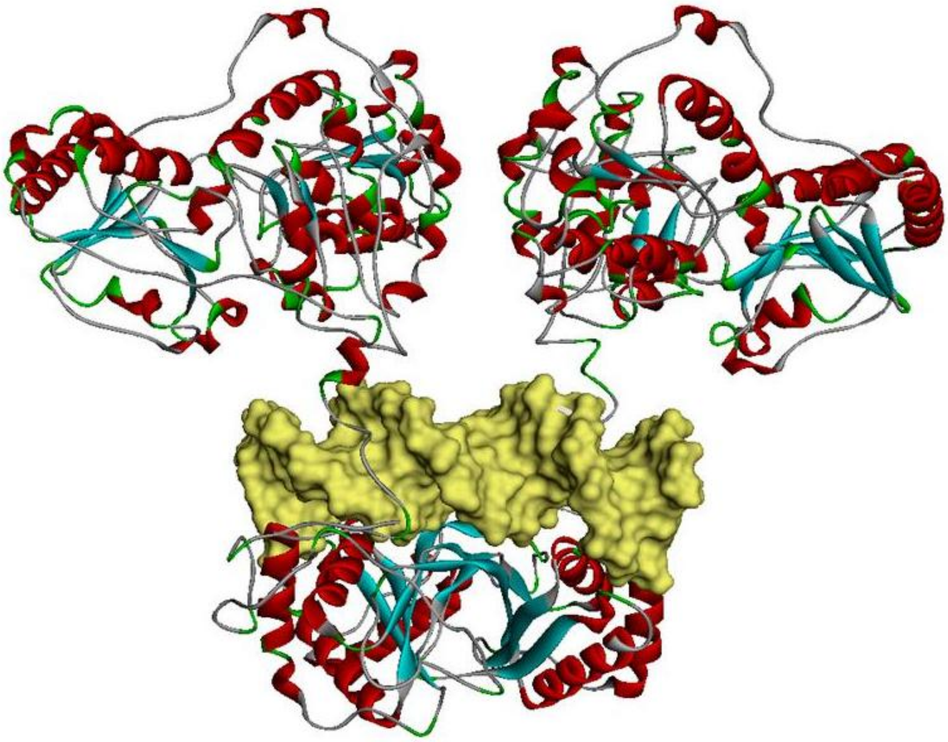


Figure 3.6. EBV EBNA1 MOE dimer model. EBNA1 full length dimer as modelled in MOE is shown in ribbon format. The DNA binding region is shown in space filled topology. Note the string of residues sits in the major groove and connecting the N-terminal and C-terminal regions of the protein.

asymmetry of the monomers might result from the chance selection of different short PDB fragments for constructing the outgroup (N-terminal and C-terminal ends) modelling. Alternatively, the asymmetrical nature of DNA produce the asymmetry in the bound proteins modelled.

In the second approach, an EBNA1 dimer was constructed using SymmDock providing the composite monomer model and the intermolecular contact points, as described for the resolved EBNA1 C-terminal dimer, as input. The top 20 predictions of SymmDock were evaluated and the best model was selected on the basis of lowest free energy. The selected model shows high similarity in the orientation of the DNA recognition and binding region to the resolved C-terminal domain dimer (Figure 3.8). The N-terminal portions of both monomers orient roughly perpendicular to the C-terminal DNA binding domain. The C-terminal domains of the monomers of the SymmDock predicted dimer prediction are at a slightly altered angle to one another in comparison to 1B3T, resulting in a subtle widening of the β barrel (as compared to 1B3T) which is formed at the interface of the core domains

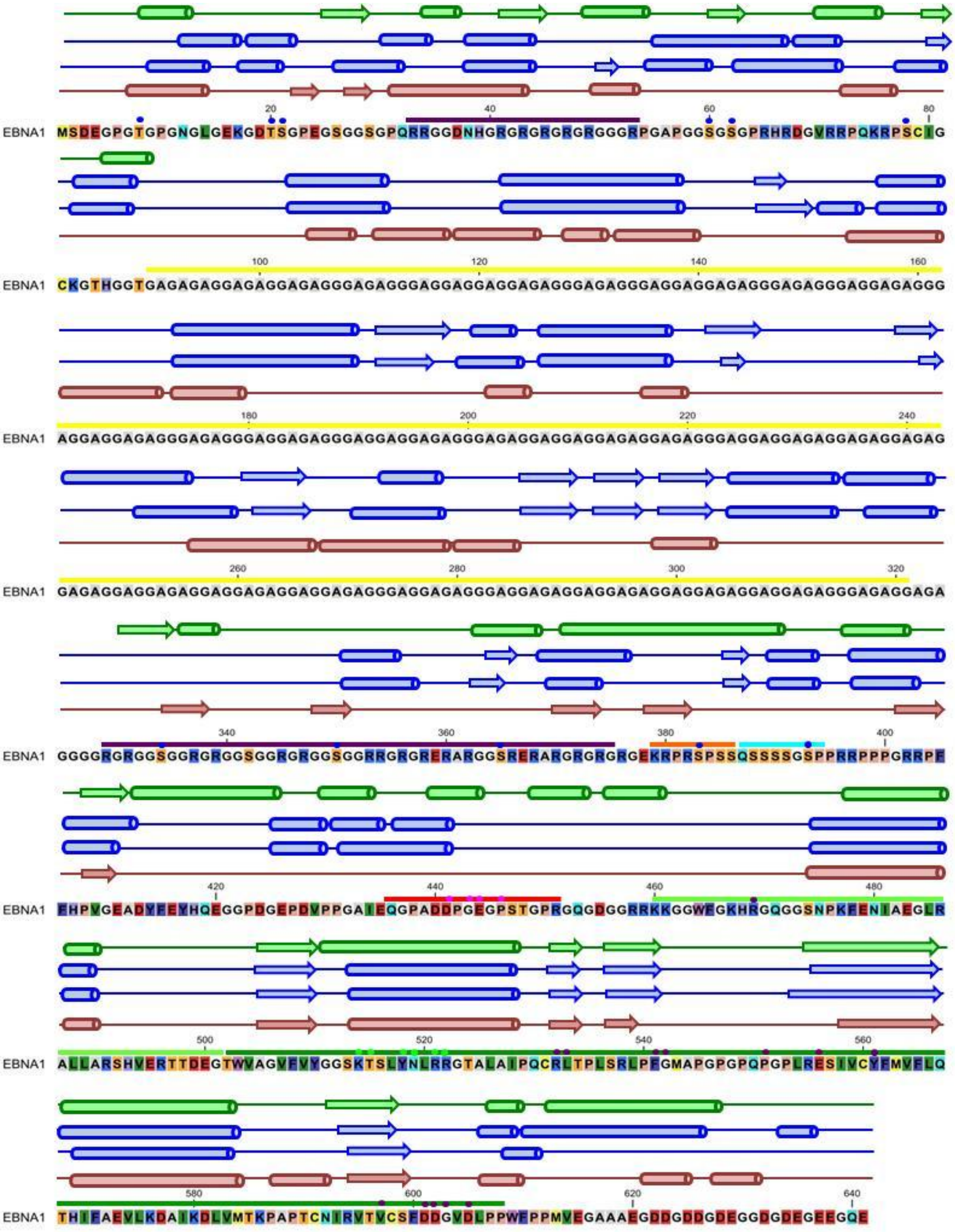

20
40
60
80
EBNA1 MSDEGPGTGPGNGLGEKGDTSGPEGSGGSGPQRRGGDNHGRGRGRGRGRGGGRPGAPGGSGSGPRHRDGVRRPQKRPSCIG
100
120
140
160
EBNA1 CKGTHGGTGAGAGAGGAGAGGAGAGGGAGAGGGAGGAGGAGGAGAGGGAGAGGGAGGAGGAGAGGGAGAGGGAGGAGAGGG
180
200
220
240
EBNA1 AGGAGGAGAGGGAGAGGGAGGAGAGGGAGGAGGAGAGGGAGAGGAGGAGGAGAGGAGAGGGAGGAGGAGAGGAGAGGAGAG
260
280
300
320
EBNA1 GAGAGGAGGAGAGGAGGAGAGGAGGAGAGGGAGGAGAGGGAGGAGAGGAGGAGAGGAGGAGAGGAGGAGAGGGAGAGGAGA
340
360
380
400
EBNA1 GGGGRGRGGSGGRGRGGSGGRGRGGSGGRRGRGRERARGGSRERARGRGRGRGEKRPRSPSSQSSSSGSPPRRPPPGRRPF
420
440
460
480
EBNA1 FHPVGEADYFEYHQEGGPDGEPDVPPGAIEQGPADDPGEGPSTGPRGQGDGGRRKKGGWFGKHRGQGGSNPKFENIAEGLR
500
520
540
560
EBNA1 ALLARSHVERTTDEGTWVAGVFVYGGSKTSLYNLRRGTALAIPQCRLTPLSRLPFGMAPGPGPQPGPLRESIVCYFMVFLQ
580
600
620
640
EBNA1 THIFAEVLKDAIKDLVMTKPAPTCNIRVTVCSFDDGVDLPPWFPPMVEGAAAEGDDGDDGDEGGDGDEGEEGQE

Figure 3.7. Distribution of structural elements on hu-EBNA1 models. The distribution of structural elements on the hu-EBNA1 as observed in the *in silico* models (maroon (composite model), blue (MOE models) and green (composite GAr deleted model)) are shown as cylinders (α helices) and arrows (β sheets).

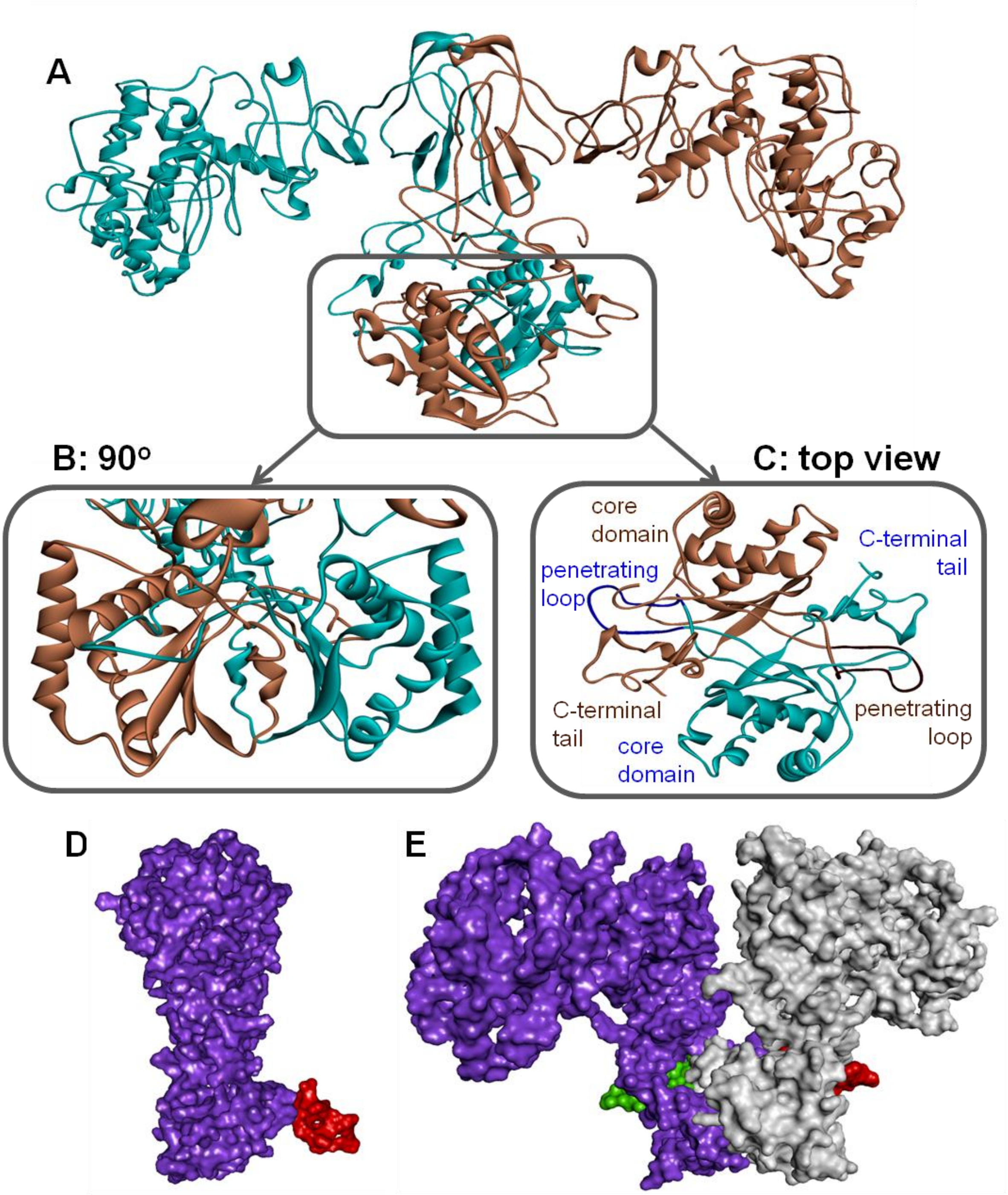


Figure 3.8. EBV EBNA1 composite model dimer structure. (A) EBV B95-8 EBNA1 composite dimer model is shown in the ribbon cartoon with the each monomer coloured brown or cyan. The C-terminal region associated with the DNA binding and homodimerisation is highlighted in a box, magnified and rotated by 90^{o} in the horizontal plane to show the central barrel **(B)** and viewed from the top to visualize protrusion of proline loop **(C)**. Space filled topology of the monomers and dimers are shown in (**D** and **E**), illustrating the proline loop of each monomer: coloured green and red, corresponding to grey and purple monomers respectively.

of both monomers. The selected model shows noticeable congruencies with the resolved EBNA1 C-terminal dimer (1B3T) in the hydrogen bond pattern and over all topology (RMSD value 1.50Å) (Figure 3.8 and 3.9). Five hydrogen bonds (Glu367-Arg368 x 2, Arg370-Ser368, Arg382-Asp455, Ser386-Asp455) lie between N-terminal residues in the SymmDock predicted dimer are not present in the resolved C-terminal DNA binding and dimerisation domain. Nine hydrogen bonds (Tyr510-Trp609, Tyr510-Phe610, Arg532-Gly542, Leu533-Pro553, Ile558-Glu629 x 2, Tyr561-Trp609, Arg594-Pro608, Arg594-Phe610) are located in the C-terminal region of the predicted dimer (Figure 3.9). The counter-part of some intermolecular hydrogen bond contacts differ between the resolved structure (1B3T) and the SymmDock dimer, for example Arg594-Asp605 bond found in 1B3T is replaced with Arg594-Pro608 in the predicted dimer but of note, Pro608 was not present in the crystallized protein-DNA complex, therefore it is possible this contact in the 1B3T may be artifactual.

Interestingly, the protruding proline loop of one monomer slots into the hole between the C-terminal tail and the core domain of the other monomer (Figure 3.8). In the resolved structure the C-terminal tail (residues: 608-641) was not included thereby the significance of this conformation was not apparent. Although the C-terminal tail of the dimer predicted by MOE still encircles the proline loop to some extent it does not exhibit a complete ring. In total, based on the orientation of the proline rich loop in the predicted EBNA1 dimers, we propose that the protrusion of the proline rich loop of a monomer into another of its counter-part may render stability in the EBNA1 dimerisation. To evaluate this hypothesis, another dimer was constructed using SymmDock, using a monomer model deleted for the C-terminal region in question (deleting residues: 608-641). Of the top 20 best predictions, only 8 show the correct orientation of the two monomers to each other (in relation to the resolved C-terminal structure) as compared to 15/20 for the full length EBNA1 dimer predictions. Additionally, the best dimer shows poor values for geometry and free energy compared to the full length dimer, supporting the hypothesis that the C-terminal tail may be involved in dimer stabilisation.

In comparison to the MOE predicted monomer and dimer, the composite model of EBNA1 and SymmDock generated dimer give improved structural reliability scores. However, the SymmDock dimer shows spatial constraints at the location where DNA should interact with the C-terminal domain of the protein. Structurally, this region is predicted to be disordered in both models and FoldIndex analysis. Therefore it is likely that this region may be flexible and adjust its position during DNA interaction.

Deletion of GAr of hu-EBNA1 facilitates increased expression of the protein in heterelogous systems, while retaining several of its functions including the ability to bind

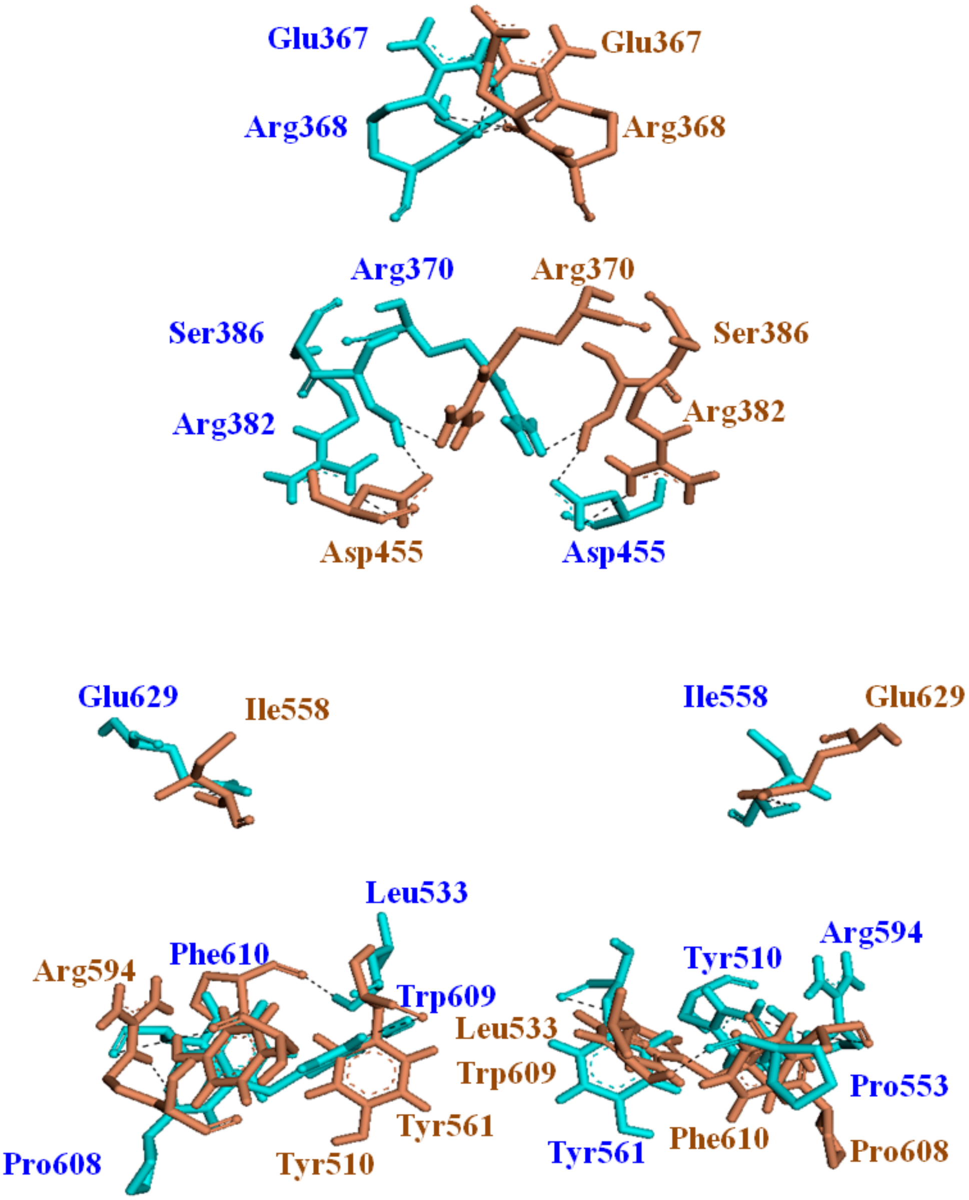


Figure 3.9. Hydrogen bonds in EBNA1 composite dimer model. The intermolecular hydrogen bond pattern between the two EBNA1 monomers (cyan and brown) is shown. Residues involved in the bond formation are labelled with their respective positioning and hydrogen bonds are represented by black dashed line.

with DNA and genome maintenance. A composite GAr deleted model was generated by employing the same methodology as used in the construction of the full length EBNA1 composite model. Subsequently the GAr deleted dimer was constructed using SymmDock. Consistent with the full length EBNA1 model, the GAr deleted dimer model also retains the same conformation of C-terminal DNA binding and dimerisation domain as found in 1B3T. However, the C-terminal tail in the GAr deleted EBNA1 model is predicted to be structured and composed of a long α helix which forms a half ring (Figure 3.10). Exclusion of GAr in the model impacts the protein interacting region as both CK2 and USP7 binding regions forms short alpha helices. Interestingly, three hydrogen bonds that are found in the 1B3T structure and absent or differently paired in the full length EBNA1 model (e.g. Arg469-Glu556) are present in the GAr deleted EBNA1 dimer (Table 3.2).

Dimer models were also constructed from the EBNA1 monomers of non-human primate LCV using SymmDock (Figure 3.11). Given the regional distribution of sequence conservation, as expected in all cases the C-terminal domain participates in the dimerisation while the N-terminal half orients approximately perpendicular to the C-terminal domain. However, some important differences were noted between the hu-EBNA1 dimer and the other primate EBNA1 dimers, for example the β barrel of Old World monkey LCV EBNA1 structures is wider compared to hu-EBNA1. By contrast the β barrel of ma-EBNA1 dimer is similar to the hu-EBNA1 dimer in terms of symmetry of the interacting interface of the monomers. In summary, the modelled structures predict that homodimerisation of EBNA1 is a conserved structural characteristic that is retained in all primate LCV EBNA1 molecules.

3.6. Zinc binding with EBNA1 homodimer

Two conserved Cys residues in LR1 (Cys79 and Cys82) have been found to coordinate zinc to potentially facilitate interaction (linking) between EBNA1 dimers to form a homo-multimeric complex at the repeated binding sites (the family of repeats, FR) at *oriP* (Aras *et al*., 2009). To explore this structurally, two zinc ions were introduced in the EBNA1 dimer using MOE. Energy minimization (Amber forcefield 99) predicted the stable bonding between these two cysteines and the zinc ions (Figure 3.12), supporting this mode of linkage between the adjacent EBNA1 dimers while interacting with FR.

3.7. Phosphorylation in EBNA1

Several post translation modifications of the EBNA1 molecule have been reported, including phosphorylation of serine residues and methylation of arginine residues (Shire *et al*., 2006; Duellman *et al*., 2009). In order to explore this structurally, the composite model of EBNA1 was examined for the propensity of serine residues to undergo phosphorylation.

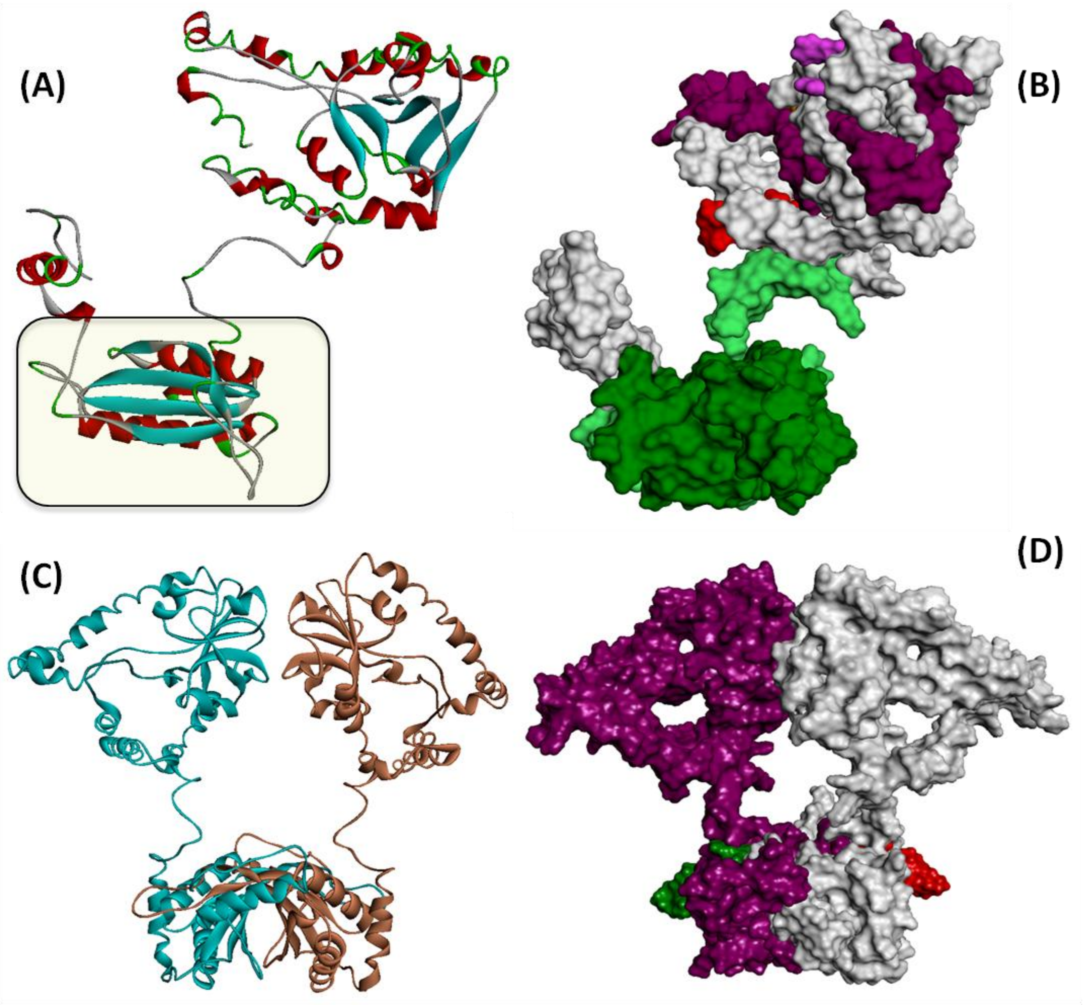


Figure 3.10. GAr deleted EBNA1 monomer and homodimer. The composite model of GAr deleted EBNA1 is shown in both ribbon format **(A)** and surface view **(B)**. The boxed region **(A)** indicates the region which has been structurally resolved (in 1B3T). The surface topology images are colour coded to highlight structural and/or functional yellow: GAr; purple: GR1 and GR2; pink: Arg71 and Arg72; cyan: CK2 interaction region; orange: NLS; red: USP7 binding site; light green and dark green: flanking region and core DNA binding and dimerisation domain. The GAr deleted EBNA1 monomer model (**A**, **B**) was used to generate a dimer in SymmDock (**C**, **D**). Ribbon format views are shown with each monomer coloured cyan and brown (**C**). Monomers/proline rich loops in the surface topology view (**D**) are differently coloured: mauve/red and silver/green corresponding to purple and grey coloured monomer respectively.

No.	1B3T	Full length model	Gly/Ala deleted model
1	NS	Not found	Lys313-Arg314 (x2)
2	NS	Glu367-Arg368 (x2)	Not found
3	NS	Arg370-Ser386	Not found
4	NS	Arg382-Asp455	Not found
5	NS	Ser386-Asp455	Not found
6	Arg469-Glu556 (x2)	Not found	Arg469-Glu556
7	Tyr510-Asp605	Tyr510-Trp609	Tyr510-Asp605
8	NS	Tyr510-Phe610	Not found
9	NF	Not found	Arg521-Pro553 (x2)
10	Arg532-Gly542 (x2)	Arg532-Gly542 (x2)	Arg532-Met543 (x3)
11	Arg532-Phe541	Not found	Arg532-Gln550
12	Leu533-Pro553	Leu533-Pro553	Not found
13	Gly542-Pro607	Not found	Not found
14	NF	Not found	Ala544- Glu641
15	NF	Not found	Arg555-Gly470 (x2)
16	NF	Ile558-Glu629 (x2)	Not found
17	Tyr561-Tyr561	Tyr561-Trp609	Tyr561-Tyr561
18	NF	Not found	Ala588-Asp625
19	NF	Not found	Cys591-Asp625
20	Arg594-Asp605	Arg594-Pro608	Not found
21	NS	Arg594-Phe610	Not found
22	Thr596-Asp602	Not found	Not found
23	Val597-Asp601	Not found	Not found
24	Ser599-Ser599	Not found	Not found

Table 3.2. Hydrogen bond interactions in EBNA1 dimer. Comparison between inter-residual hydrogen bonds of structurally resolved C-terminal EBNA1 dimer (1B3T) (residue 461-607), composite full length EBNA1 dimer (residues 1-641) and GAr deleted EBNA1 dimer (1-90; 327-641) are tabulated. Note, in the GAr deleted EBNA1 dimer, amino acids are numbered according to the full length protein.

Out of the total 27 serine residues, present in EBNA1 of EBV-B95-8, 13 serines were predicted for their potential to undergo phosphorylation in the given composite model (Figure 3.13). Of these, only 3 serine residues (Ser60, Ser62 and Ser393) are consistent with the 10 reported phosphorylation sites (Duellman *et al.*, 2009). It is important to note that most of the reported predicted sites are situated in the intrinsically disordered region of the molecule. Therefore, it is conceivable that the difference between the observed and predicted (structurally) phosphorylated sites could be due to the spatial restraints in the predicted model. Given the flexibility of these region it is possible that other sites (especially those of observed) will get phosphorylated.

3.8. Structure of proposed Dimer Inhibitory Peptide (DIP)

It was observed in all dimer predictions (described in section 3.5) that a protruding proline rich loop of each monomer slots into a hole (formed between the C-terminal tail and DNA

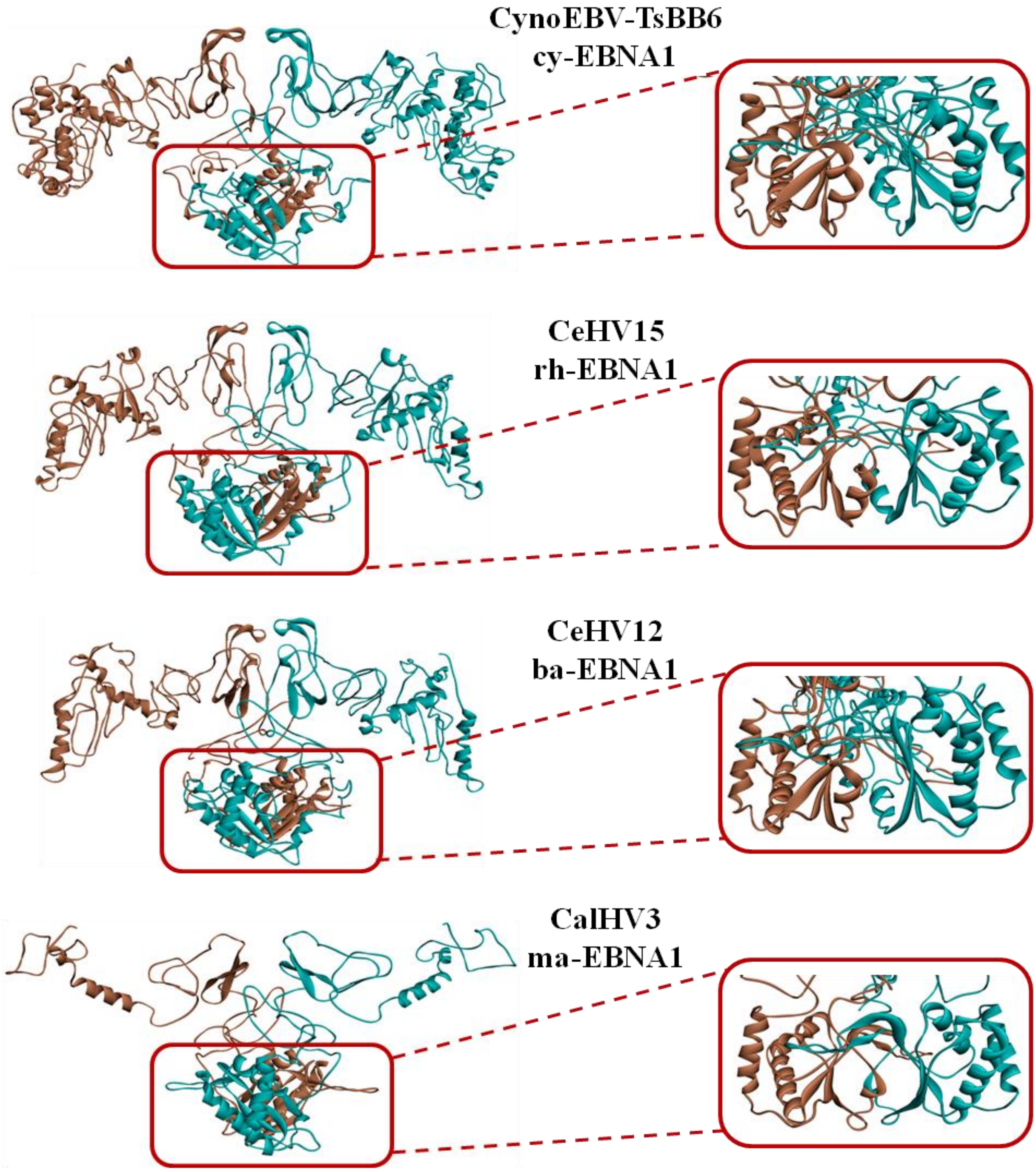


Figure 3.11. Primate LCVs EBNA1 dimers. Homodimers were constructed using SymmDock for each of the modelled (non-human) primate LCV EBNA1 homologues (as indicated). Monomers are coloured cyan or brown. Consistent with the hu-EBNA1, all primates LCVs EBNA1 dimerisation is mainly mediated by the C-terminal domain (boxed). The C-terminal regions of each predicted dimer are shown enlarged and rotated by 90^{o} in the horizontal plane.

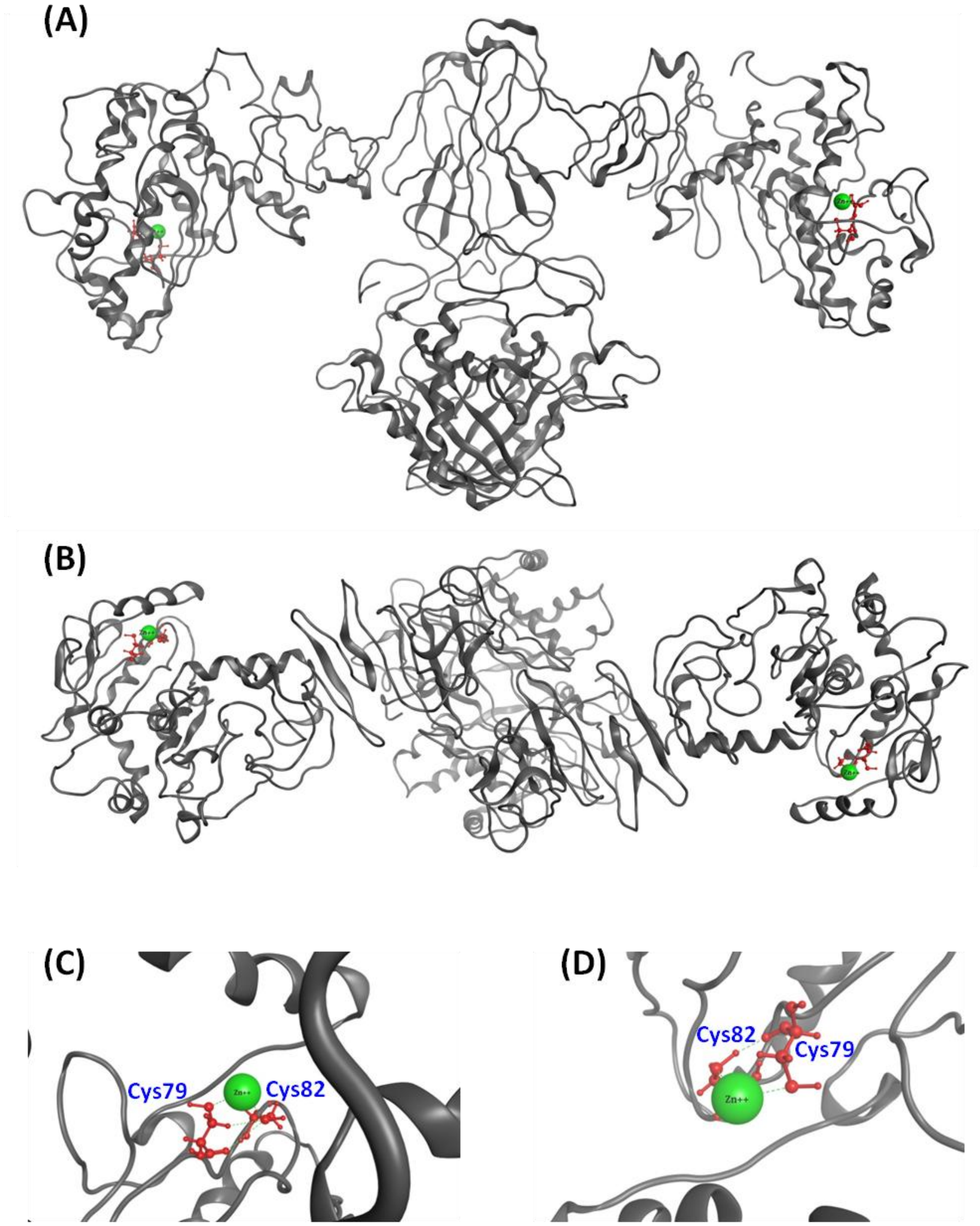


Figure 3.12. Zn incorporation in the EBNA1 dimer model. Stable binding between a zinc ion (green) and Cys79 and Cys82 (red ball and stick) of each EBNA1 monomer is predicted in the dimer model, shown in ribbon format with 90^{o} rotation in the vertical plane between **(A)** and **(B)**. **(C)** and **(D)** show enlarged images of these interactions as seen in **(B)**.

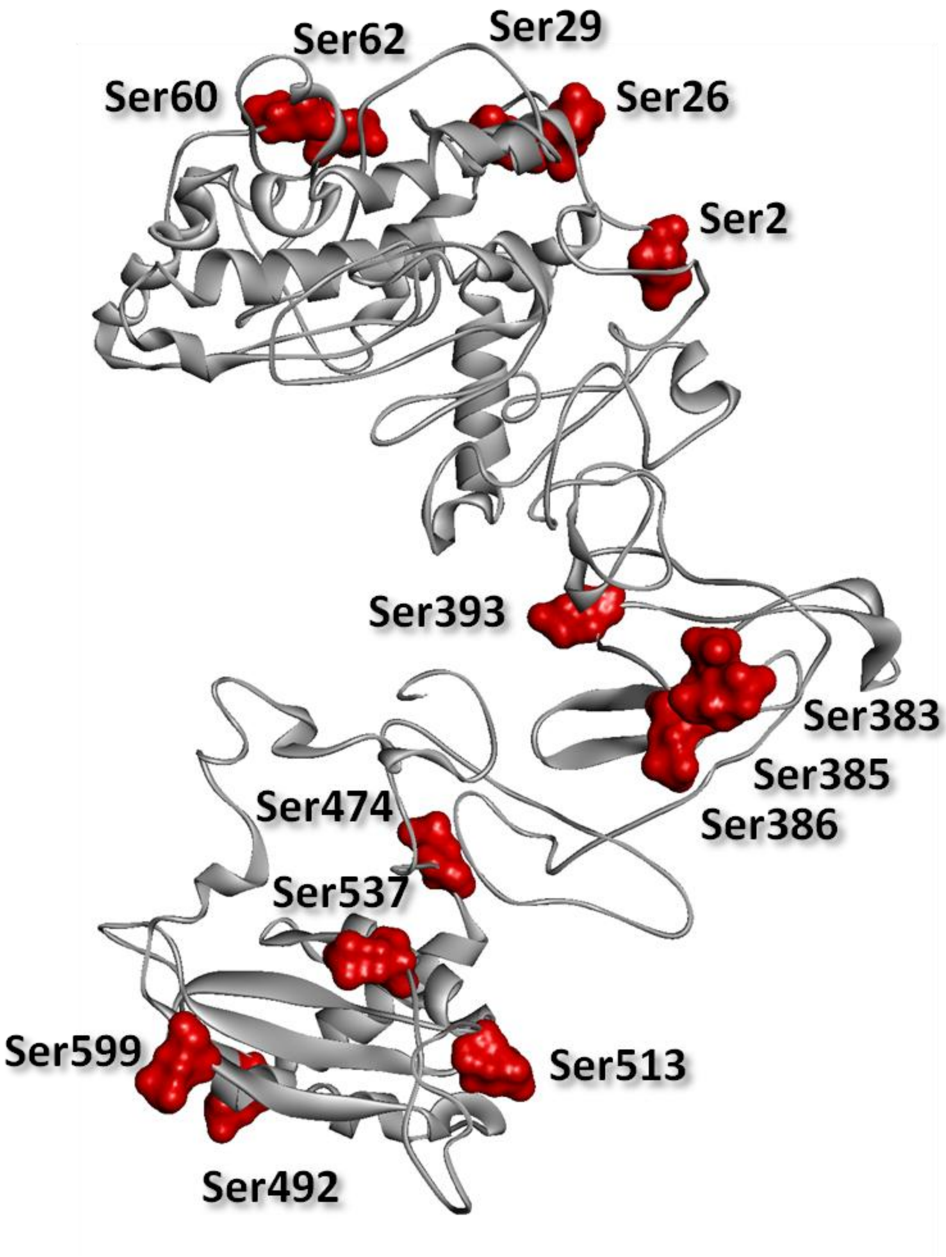


Figure 3.13. Phosphorylation of EBNA1. Ribbon cartoon of EBNA1 protein model (composite) is shown where serine residues with a propensity to undergo phosphorylation are highlighted in red surface format.

binding region) of the other monomer. This consistent positioning of the proline loop in the proposed EBNA1 dimers may have a biological role in stabilising the dimer and could be exploited to design a disruptor. To investigate this possibility a small peptide (referred to as DIP) was designed by combining the TAT (YGRKKRRQRRR) sequence with the proline loop sequence (residues: Phe541-Arg555) at the N-terminal. The TAT sequence, found in the human immunodeficiency virus (HIV) TAT protein, is known for its ability to transport itself and linked sequence through cell and nuclear membrane (Ziegler *et al.*, 2005). As EBNA1 is a nuclear protein, therefore peptide/disruptor designed to disrupt any EBNA1 function (such as dimerisation) through direct physical interaction must cross the nucleic membrane barrier. Thus the TAT sequence was incorporated to drive the fusion peptide into the nucleus. The full peptide of 26 amino acids was modelled using I-TASSER. In the *in silico* model of DIP, the N-terminal TAT sequence is predicted to form a helix while the proline loop shows a disordered conformation, consistent with that observed all EBNA1 monomer and dimer models (Figure 3.14). However, the intramolecular hydrogen bonds of the proline loop between DIP and the full length EBNA1 molecule differ, the number of predicted intramolecular hydrogen bonds were reduced from 7 (proline loop in native conformation) to 2 (proline loop in DIP). Importantly, 2 of 7 of these bonds in the proline loop of the full length EBNA1 model are present at turns (Gly548-Gly546; Gly546-Ala544) or between two limbs of the loop (Gln550-Met543). In contrast both hydrogen bonds (Gly542-Arg11; Leu554-Gly522) in the proline loop of the DIP model are formed at its terminal ends. It is possible that incorporation of the TAT peptide may influence the intramolecular interactions and topology of the proline loop in the DIP molecule or that taken out of context, it loses its configuration.

3.9. Molecular docking of DIP with EBNA1 monomer

To explore the potential binding orientation of DIP with the EBNA1 composite monomer, the DIP model was docked (blind i.e. without inputting required contacts) against the composite EBNA1 monomer model using ClusPro2.0 (Comeau *et al.*, 2004). The top 10 ranked simulations under balanced coefficient were examined for the position of DIP on the EBNA1 model and are collectively illustrated in Figure 3.15. The best or lowest docking score of balanced interaction mode is -1225.284 Kcal/mol and the highest is -979 K cal/mol and the average score of docking energy center is 1081.4 Kcal/mol. In all docking simulations DIP was predicted to interact within or near to the potential target region (the central hole formed between the C-terminal tail and the DNA binding region) of the EBNA1 monomer. On the basis of orientation of the DIP to EBNA1, these

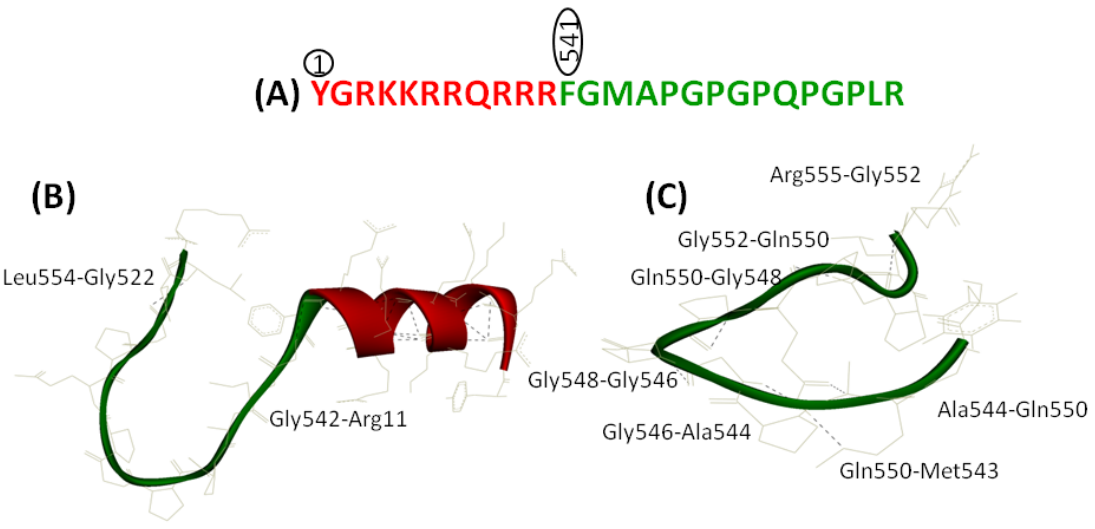


Figure 3.14. Structural models of DIP and proline loop of full length EBNA1. Protein sequence **(A)** and predicted model **(B)** of DIP are shown where TAT and proline loop are coloured red and green respectively. The residue numbering of the proline loop is according to the full length EBNA1 sequence and indicated in **(A)**. Amino acids are represented by grey lines and black dashed lines represent the hydrogen bonds between the residues indicated. Note the difference in the numbers of intramolecular hydrogen bonds in proline loop of DIP **(B)** and modelled structure of EBNA1 **(C)**.

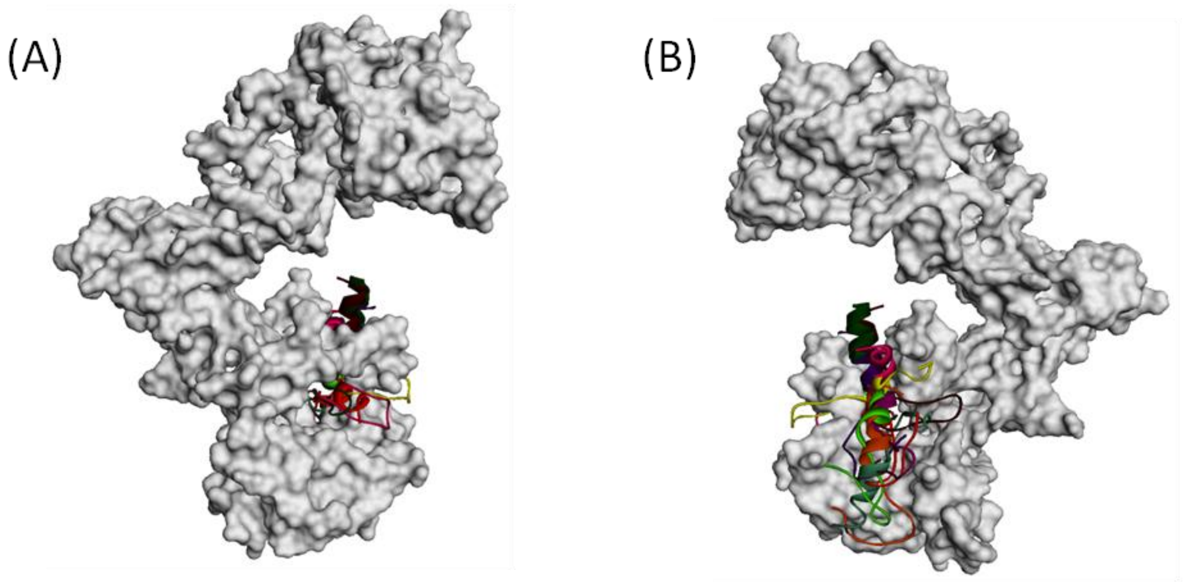


Figure 3.15. Molecular docking of DIP with EBNA1 monomer. Molecular docking simulations were conducted using ClusPro2.0 and the top 10 docking simulations, ranked according to balanced coefficient, are shown here (**A** and **B**: 180^o rotation). DIP is represented in ribbon format while the EBNA1 composite monomer is represented in surface topology with 180^o rotation in the horizontal plane between the two images. Each of the 10 spatial positionings of DIP corresponds to a single simulation which are colour coded according to the rank: 1st (purple); 2nd (red); 3rd (orange); 4th (pink); 5th (barn red); 6th (lemon); 7th (yellow); 8th (fuchsia); 9th (amazon) and 10th (green). Note: All of the top 10 ranked simulations place DIP at or within the proline loop hole of EBNA1.

interactions could be broadly classified into two different conformations: 1) penetrating or 2) capping. Capping conformations (in which the DIP covers one side of the central hole) are slightly favoured over penetrating conformations (where dimer proline loop penetrates the central hole) with 60% and 40% ratio respectively (Figure 3.15). Given the docking simulations were undertaken without specifying the target region, the orientation of DIP on EBNA1 and associated energy values are encouraging. If DIP will associate with EBNA1 in either orientation *in vivo* (penetrating or capping), it is expected to interrupt and/or hinder the protrusion of the proline loop (of another monomer) into the central hole of the 1st monomer, thereby inhibiting/destabilizing homodimer formation.

3.10. DIP uptake

Based on the *in silico* docking analysis of DIP with EBNA1, preliminary empirical study was designed to examine the effect of DIP on the EBNA1 dimerisation. First, DIP was synthesized with and without fluorescein (Flu) tag (DIP and DIP-Flu respectively). Various B cell lines (positive and negative for EBV; Table 2.5) were treated with 10µM DIP-flu and uptake and intracellular localization of the peptide were examined using confocal microscopy. After 20 minutes of peptide (DIP-Flu) treatment, the cell lines were stained with DAPI and examined by confocal microscopy (Figure 3.16). Amongst the observed fields, most cells of all cell lines have shown penetration of DIP. The DIP-Flu was mainly distributed in nucleus and in some cells accumulation of peptide in nucleoli was also noticed (Figure 3.16).

3.11. EBNA1 expression in different cell lines

In order to select the B cell line for the preliminary studies of DIP, EBNA1 expression in several B cell lines (Table 2.5) was assessed. Expression of the protein was examined by western blotting using 100µg of cell line lysate and probing with anti EBNA1 antibodies (IH4). Immuno blotting showed maximum expression of EBNA1 in Raji cell line and least expression was observed in BL30 (EBV positive) cell line. In comparison to Raji, much less expression of EBNA1 was observed in IB4, B958 and Namalwa (in decreasing order) (data not shown). In a separate experiment, EBNA1 expression was also assessed by comparing the lysates of 10^6 cells of Raji, Namalwa and as negative control BL2 (EBV negative) (Figure 3.17). Higher expression of EBNA1 was observed in Raji cell lysate followed by Namalwa while BL2 (EBV negative) showed no band of protein at the corresponding size of EBNA1 when probed with IH4 antibodies (Figure 3.17). This is consistent with the earlier studies demonstrating the presence of only 2 integrated copies of EBV genome in each Namalwa cell (Ryan *et al*., 2004) compared to 50 to 60 EBV genome copies per Raji cell (Adams *et al*., 1973). Taking both cellular penetrations of DIP

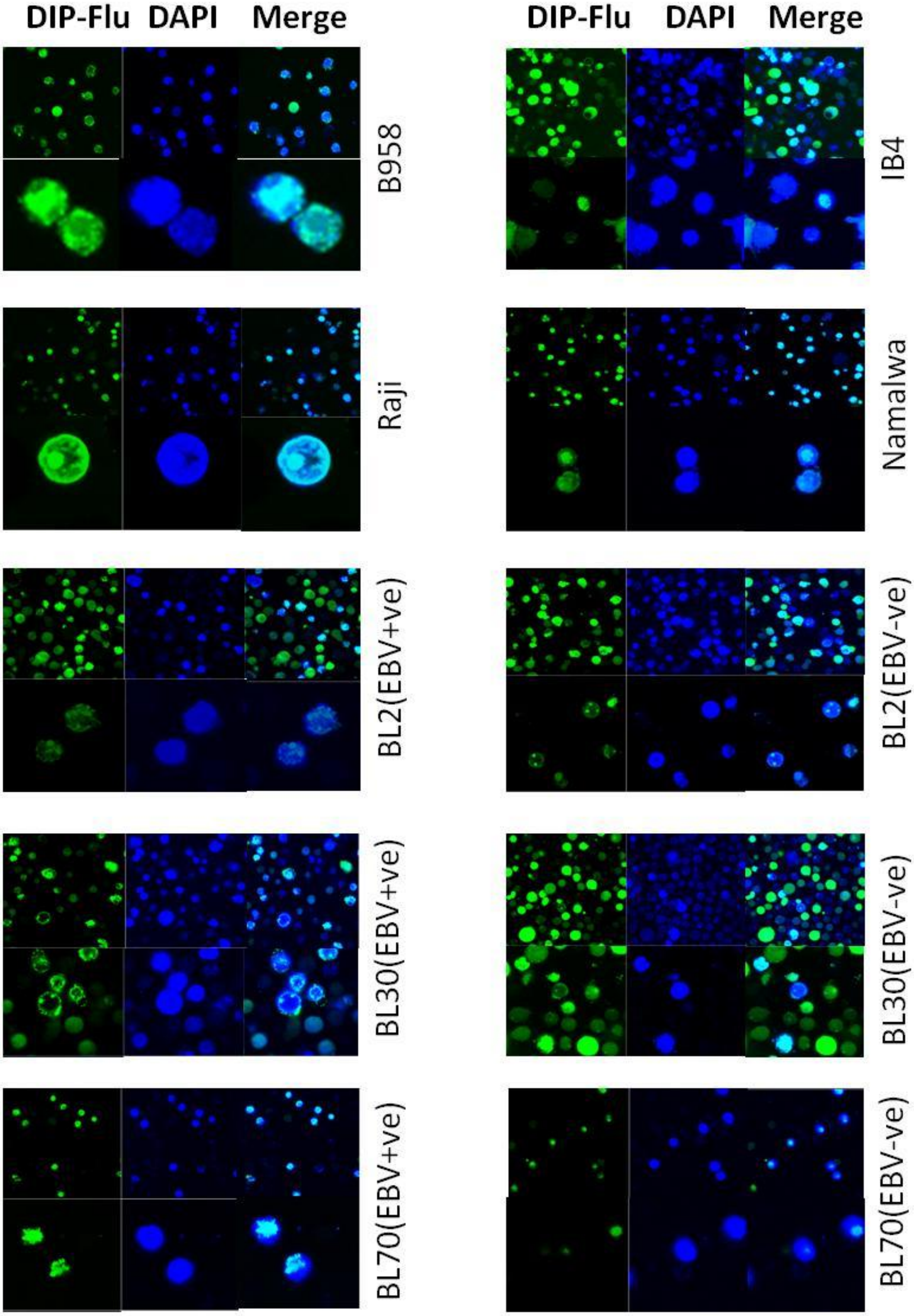


Figure 3.16. Uptake of DIP-Flu in B cell lines. Several cell lines (as indicated) were treated with 10µM DIP-Flu peptide for 20 minutes and the cells were counter stained with DAPI. Cells were observed without fixation in different planes and fields of view. Images are representative of one field of view observed and were taken at different magnifications (for each cell line; top panel images (100x) and bottom panel images (600x) except Raji and B958 images were taken at 1200x) and with different contrasts as indicated.

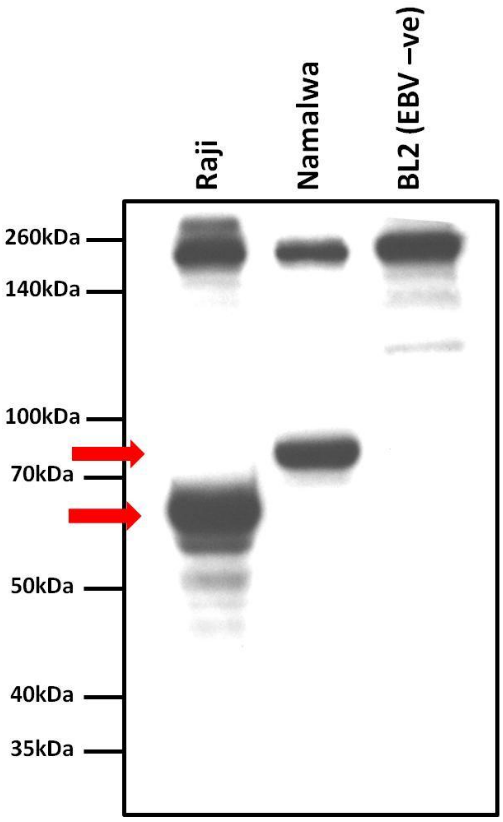


Figure 3.17. EBNA1 expression in B cell lines. Cell lysate (10^6 cells) of different B cell lines (as indicated) were probed with anti EBNA1 antibodies (IH4). The expected size of EBNA1 protein are indicated with red arrows. Note: the difference in the molecular weight of EBNA1 between Raji and Namalwa cell lines is due to the deletion of GAr domain in the former.

and EBNA1 protein expression, Namalwa cells seems an adequate system for preliminary examination of the effect of DIP on the EBV infected cell.

3.12. DIP inhibits cellular proliferation of Namalwa cell line.

In order to explore that DIP inhibits the growth of the EBV dependent cell line, via disrupting essential EBNA1 function(s), Namalwa cell line was treated with 25µM DIP. To rule out any observed effect due to the cellular toxicity of the DIP, BL2 (EBV negative) cell line was also treated with the same concentration of DIP. The experiment was conducted in triplicate for four days and on each day cells were counted by trypan blue exclusion assay (Figure 3.18). In another set of control, Namalwa and BL2 cell lines were grown in the absence of DIP. During the first two days, decline in the cell numbers were observed in all cell lines however, in the next 3 days, all cell lines recovered and started to proliferate, except DIP treated Namalwa cell line. No difference in the growth was

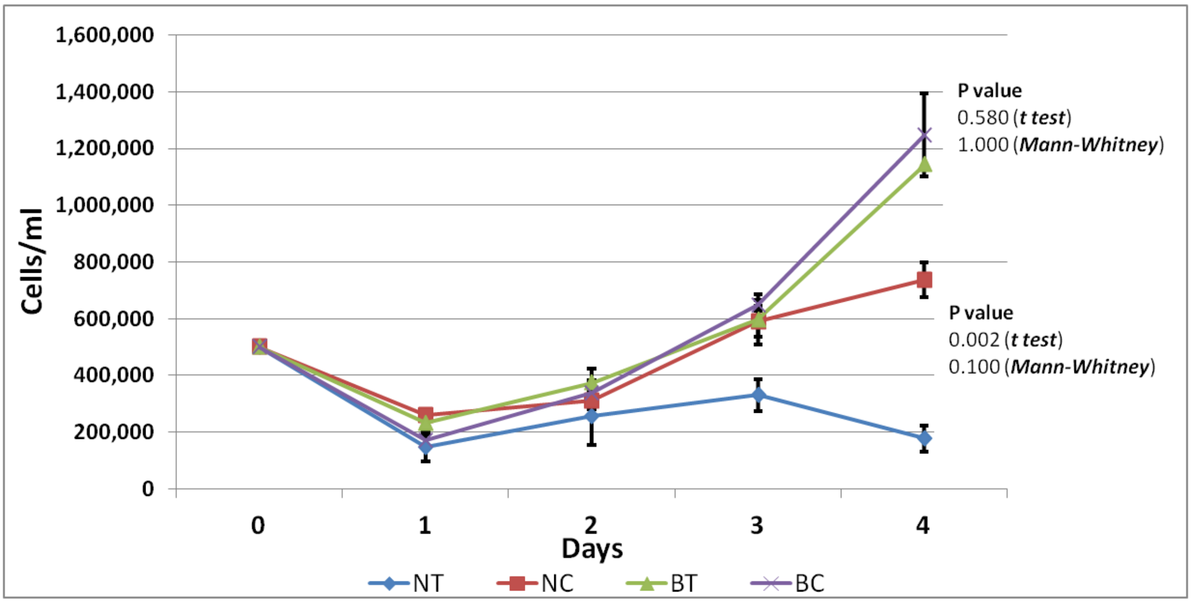


Figure 3.18. DIP effect on the EBV positive B cell growth. Namalwa and BL2 cells (10^6) were treated 25µM of the DIP. Cells were counted over a four day period by trypan blue exclusion assay to assess the cellular proliferation. Each data point reflects the mean count of three replicates with vertical bars represent standard error of mean. **Key:** Namalwa cells treated with DIP (blue line; NT), Namalwa cells control (maroon line; NC), BL2 cells treated with DIP (green line; BT) and BL2 cells control (purple line; BC)

observed between DIP treated and untreated BL2 cell lines. However, relatively slow growth was observed in DIP treated Namalwa cells compared to its respective negative control. On the final day (day 4) in the DIP treated Namalwa, most cells were dead while the untreated Namalwa still showed increase in the number of cells and the difference in the final count of cells in both (DIP treated and untreated Namalwa cells) was found to be statistically significant under student t test but not when Mann-Whitney test was employed. In total the data suggest that DIP may retard the proliferation of EBV positive Namalwa cell line and the continuous treatment of DIP led to the death of the Namalwa cells. As mentioned that this experiment is a preliminary investigation for evaluating the effect of DIP in EBV positive cell line and essentially require further studies (discussed latter) for further verification.

3.13. Summary of findings

- Marmoset LCV EBNA1 lacks several functional regions found in human EBV EBNA1, there are a split GR domain, the entire GAr domain and CK2 and USP7 binding sites.
- The GAr region shows a sequential and structural increase in length in the EBNA1 of Old World Monkey LCVs from baboon to humans.

- The structural model of EBNA1 shows a strong alignment with the resolved C-terminal DNA binding and dimerisation region while the N-terminal region is largely composed of helices and unstructured loops.
- The protein interaction sites in EBNA1are largely comprised of intrinsically disordered regions, which may allow EBNA1 to bind with multiple alternative partners.
- The C-terminal tail of EBNA1 is composed of an unfolded loop which curls back to the DNA binding and dimerisation domain of the protein providing a hole for insertion of the loop of the interacting EBNA1 in the dimer.
- EBNA1 models of primate LCVs are similar in over all topology and domain architecture (where present) to EBNA1 from human EBV.
- It was possible to generate dimer models in SymmDock or MOE from the full length monomer models without any spatial restraints and intermolecular clashes.
- Unlike the EBNA1 dimer constructed using MOE, a dimer model constructed using SymmDock showed spatial interference with a space likely to be occupied by DNA due to the occupancy of the loop constituting the partner proteins (EBP2, CK2 and USP7) binding sites. However, the dynamic nature of protein structure may allow flexibility in this region allowing it to change its spatial position and/or structure upon the interaction of EBNA1 with partner molecules, including DNA.
- A plausible coordination between Zn^{2+} and two conserved cysteine residues (towards the N-terminus) was observed using the EBNA1 model, consistent with the multimeric complex formed by EBNA1.
- We have hypothesized that the GAr region may mask the unstructured region of the EBNA1 protein to prevent its proteolytic degradation and consequent epitopic presentation.
- In all EBNA1 dimer models generated, a conserved proline loop of one monomer protrudes into the space between the C-terminal tail and DNA binding and dimerisation domain of the other monomer, generating a dowel pin like joint. It is hypothesized herein that this joint may be involved in the stabilization of dimer conformation of EBNA1 and could be used as a therapeutic target.
- A disruptor peptide (DIP) has been designed for this dowel pin like joint and blind molecular docking studies indicate that DIP may interact with the EBNA1 monomer in, or near the desired region.
- Confocal microscopy experiments using fluorescent DIP showed good penetration of the peptide in all the B cell lines examined (both EBV+ve and EBV-ve).

- A cell proliferation assay showed growth inhibition in the EBV positive cell line (Namalwa) treated with DIP unlike an EBV-ve cell line, suggesting that DIP could potentially inhibit the growth of at least those EBV infected cells where survival depends upon the EBNA1 function.

3.14. Discussion

Comparison of EBNA1 sequences of primate LCVs were made to explore the evolutionary history of this viral gene. Additionally, *in silico* models of monomer and dimer conformation of full length of EBNA1 molecules of primate LCVs were generated and compared. Subsequently a possible dimerisation disruptor peptide was predicted by *in silico* means and preliminary empirical studies were conducted to evaluate these predictions.

Evolution and protein domain divergence in EBNA1 of primate LCVs

The phylogenetic analysis of EBNA1 homologues from related LCVs is in line with the evolutionary history of these viruses as inferred from the earlier analysis of DNA polymerase and glycoprotein B genes (Ehlers *et al*., 2010). It has been proposed that LCVs co-evolved with their respective host species (mostly primates) with some evidence of inter species transfer (Ehlers *et al*., 2010; Perelman *et al*., 2011). The notable differences between ma-EBNA1 and hu-EBNA1 (outside the C-terminal DNA binding and dimerisation region) suggest sub-neo-functionalization occurred in the gene after the Old and New World primates split (43 MYA). Alternatively, the similarity between hu-EBNA1 and other primates LCV EBNA1 is possibly an indicative of interspecies transfer.

Owing to the strong conservation between the C-terminal domains of EBNA1 of primate LCVs, it is likely that there is mechanistic conservation also in homodimerisation and sequence specific DNA binding. Similarly, the LR1 and GR2 regions are also highly conserved in EBNA1 homologues of Old World primate LCVs. Unlike other EBNA1 homologues, ma-EBNA1 has only one GR region, which shows similarity with GR2 in hu-EBNA1. In the absence of a GAr region, this single GR region of ma-EBNA1 is located near the N-terminus of the protein. It is possible that this single GR domain represents the ancestral form of GR regions prior to the incorporation/emergence of the GAr domain in EBNA1. Deletion of LR1 from hu-EBNA1 adversely affects its transactivation function (Singh *et al*., 2009) and therefore ma-EBNA1 may differ with hu-EBNA1 in this respect. Both GR1 and GR2 regions of hu-EBNA1 are demonstrated to be involved in binding EBP2 and G-rich RNA and are required for stimulation of EBNA1 dependent viral genome replication, tethering to metaphase chromosomes and faithful segregation of viral genomes in the dividing infected cells (Shire *et al*., 1999; Norseen *et al*., 2009; Shire *et al*., 2006). Sequence and structural similarities within these regions is indicative of functional conservation in viral genome propagation and segregation. A distinct NLS is absent from ma-EBNA1, however, it is possible that other regions with similarity to consensus NLS act in nuclear localization of ma-EBNA1.

The USP7 binding site is highly conserved in the EBNA1 homologues of Old World monkey LCVs, and it seems likely that they interact with host USP7. The potential CK2 binding site in Old World monkey LCVs EBNA1 is relatively long compared with hu-EBNA1. This could suggest that the EBNA1 of non-human Old World monkey LCVs may interact with CK2 with different affinity. ma-EBNA1 lacks the GAr domain, USP7 binding site and possibly CK2 binding region as well. With the availability of more EBNA1 sequences it would become clear whether or not these sites were acquired or lost together. Similar to EBV in humans, CalHV3 in marmoset is commonly associated with lymphoma. However, it can only immortalize marmoset B-cells after prolonged culture (Jenson *et al*., 2002). Nevertheless, multiple differences in the genomes are present between CalHV3 and EBV, such as the polyproline tract of EBNA2, EBERs and viral IL10 are completely absent in CalHV3 (Rivailler *et al*., 2002).

EBNA1 dimerisation and mutlimerisation in models

In all of the compared EBNA1 models, a short proline rich string within the C-terminal domain is strongly conserved and forms a protruding loop, consistent with the resolved C-terminal domain of hu-EBNA1. Interestingly, in all of the dimer models this region of each monomer inserts into the space between the C-terminal tail and core domain of other monomer, forming a “dowel pin joint” like interlocking structure. Additionally, a dimer model constructed using hu-EBNA1 monomer model with the C-terminal tail deleted, shows less favourable energy values, suggesting reduced stability compared to the full length dimer. As evident from the resolved C-terminal domain dimer, the C-terminal tail is not necessary for EBNA1 dimerisation; however, consistent in all EBNA1 dimers curling of the C-terminal tail of each monomer around the proline loop of other monomer suggests it may contribute to the stabilisation of the dimer. Importantly, this information could be exploited to design therapeutic molecules to disrupt EBNA1 dimerisation and thus associated functions. If these models are good representatives of the native conformation of EBNA1, then filling the space formed by the C-terminal tail or obstructing the proline loop to orient itself in the space, may preclude EBNA1 dimer formation.

Incorporation of zinc ions into the dimer model predict stable bonding between the zinc and Cys79 and Cys82, located at the distal end of each N-terminal arm. This is consistent with the proposal of self association of EBNA1 dimers through zinc, when interacting with

FR. Furthermore, these residues along with zinc are required for cooperative transactivation (Aras *et al*., 2010).

DNA binding with EBNA1 models

The composite EBNA1 model shows good agreement with the structure propensity predictions and resolved C-terminal domain structure for EBNA1 gives improved Ramachandran plot values and QMEANnorm scores compared to other primary models. However, unlike the model generated using MOE, the composite model shows the presence of some residues in the space which should be occupied by the DNA (when bound) as observed in the co-crystal structure of the C-terminal EBNA1 dimer and DNA. Given that these residues reside within the unstructured region of the protein, they are likely to move away upon interaction with DNA, to allow DNA to bind with the EBNA1 dimer. Therefore, it is proposed here that the composite model at present provides the best prediction of full length EBNA1 dimer conformation when not bound with DNA. It is possible to speculate that while interacting with DNA through the C-terminal domain, the unstructured strings of residues twist to provide space for DNA and in turn also move the N-terminal portion of the protein to adopt an angle in relation to C-terminus as found in dimer predicted by MOE.

Protein interaction sites in EBNA1 are intrinsically disordered

The known EBNA1-protein interaction sites (with EBP2, USP7 and CK2) are predicted to be unstructured in the *in silico* models. These unstructured regions in turn allow the protein molecule to adopt different conformational states at least in part, which could favour rapid association/dissociation, promiscuity in partner protein interactions and post translational modifications (Shire *et al*., 2006; Tompa, 2011). Proteins associated with signalling or transcriptional regulatory functions tend to contain intrinsically disordered regions perhaps to facilitate a greater repertoire of partner protein interactions (Babu *et al*., 2011). Similarly, EBNA1 interacts with multiple partners and in certain cases these binding sites overlap (e.g. with EBP2 and RNA) or present in the close proximity (CK2 and USP7 binding regions). Thus, it is reasonable to suggest that EBNA1 may adopt different conformations upon interaction with different molecular partners and thereby, the predicted composite model represents the “resting shape” of the molecule that could attain different conformations upon binding with different molecules.

The tumour suppressor protein p53 is an intrinsically disordered protein (IDP) (which also binds with USP7) and acts as a transcriptional regulator. Its tight regulation is mediated by efficient proteosomal degradation (ubiquitin dependent and independent) which is critical for the health of cell (Tsvetkov *et al*., 2009). It has been suggested that IDPs are susceptible to proteolytic degradation which are mostly mediated by the 20S proteosome

(Ubiquitin independent), hence to avoid degradation these disordered regions require to be masked (Tsvetkov *et al*., 2009; Gsponer *et al*., 2008). Interestingly, despite the predicted unstructured regions, EBNA1 is highly stable in B-cells (over 30 hours) (Tellam *et al*., 2007), suggesting the possibility that these unstructured regions are being masked in some way. By contrast, cytotoxic T-cell recognition of the EBNA1 is largely mediated by the presentation of peptide, derived from the newly synthesized EBNA1 molecule (Fu *et al*., 2004). These observations seem consistent with the predicted model of EBNA1 showing disordered regions which could be susceptible to proteolytic degradation immediately after synthesis. However, the observed stability of EBNA1 in B-cells could be due to rapid intermolecular complex formation, post translational modifications or possibly due to an intramolecular masking activity. It is tempting to speculate that the GAr domain could provide a possible masking region for the intrinsically disordered portions of the protein.

The GAr domain is predicted to be structured by both FoldIndex and in the composite model. The latter predicts that the GAr region is mainly composed of α helices while the MOE constructed model predicted that the GAr domain may contain both α helices and β sheets. Earlier studies proposed that the GAr region may be composed of β sheets (Tellam *et al*., 2001), conversely, it has been demonstrated that Ala residues have a tendency to break β sheets and form α helices. Although Gly residues tend to break both secondary structure conformations (α helices and β sheets), in combination with Ala they strongly favour the α helical conformation (Fujiwara *et al*., 2012), supporting the prediction made by the composite EBNA1 model.

Structural basis of GAr domain function in EBNA1 molecule

The hu-EBNA1 GAr region has been demonstrated to render resistance to proteosomal degradation and inhibits self synthesis, which results in impaired immune responses to EBNA1 (Levitskaya *et al*., 1997; Yin *et al*., 2003). However, the length and purity of repeats has a substantial effect upon these actions. Little or no self synthesis inhibition was observed in case of rh-EBNA1 and ba-EBNA1 which contain relatively shorter and impure GAr (Tellam *et al*., 2007). Additionally, cytotoxic T-cells tend to recognize rh-EBNA1 more efficiently compared to hu-EBNA1. Furthermore, in chimeric rh-EBNA1 in which the rh-EBNA1 GAr domain is substituted with its counterpart present in hu-EBNA1, impaired translation efficiency and endogenous processing was observed, suggesting involvement of the GAr domain of hu-EBNA1 in these functions (Tellam *et al*., 2007). It has been suggested that by stalling translation, the hu-GAr region reduces the synthesis of misfolded products thereby providing less epitopes for presentation by MHC (Yin *et al*., 2003). This hypothesis could be extended on the basis of present studies. In addition to impairing self synthesis, it is also possible that the predicted structured region of hu-

EBNA1 GAr domain may mask the disordered region of EBNA1 and in turn protect it from proteosomal degradation. Several observations are consistent with this proposal; (1) retarted translation might allow the GAr domain to fold independently and prior to whole protein folding thus inhibiting proteosomal degradation; (2) GAr masking the disordered regions would continue to prevent default degradation of the disordered region of the mature EBNA1 protein; (3) GAr mediated resistance to proteolytic degradation does not supersede ubiquitin mediated degradation; (4) This hypothesis accommodates both functions of the GAr domain: inhibiting translation and protection from the proteosomal degradation, as shown by hu-EBNA1 and would explain the observed differences between the functions of the hu-EBNA1 and both rh-EBNA1 and ba-EBNA1. It is possible that impure GAr domain sequence of Old World monkey LCV EBNA1, though unable to retard self synthesis may nevertheless efficiently mask the disordered region of the protein (Levitskaya *et al*., 1997). Finally, the proposed hypothesis also seems consistent with the evolutionary co-acquisition of the structurally disordered protein interaction sites (USP7 and CK2) and possible masking region (GAr domain).

An interesting comparison could be made between LANA1 of Kaposi's sarcoma-associated virus, which has central repeats (CR), shown to be responsible for immune evasion by inhibiting MHC-I mediated peptide presentation. However, unlike EBNA1, a separate subsection of repeats are involved in retarding the translation and inhibiting the processing of its potential epitopes (Kwun *et al*., 2011).

EBNA1 dimerisation a potential drug target

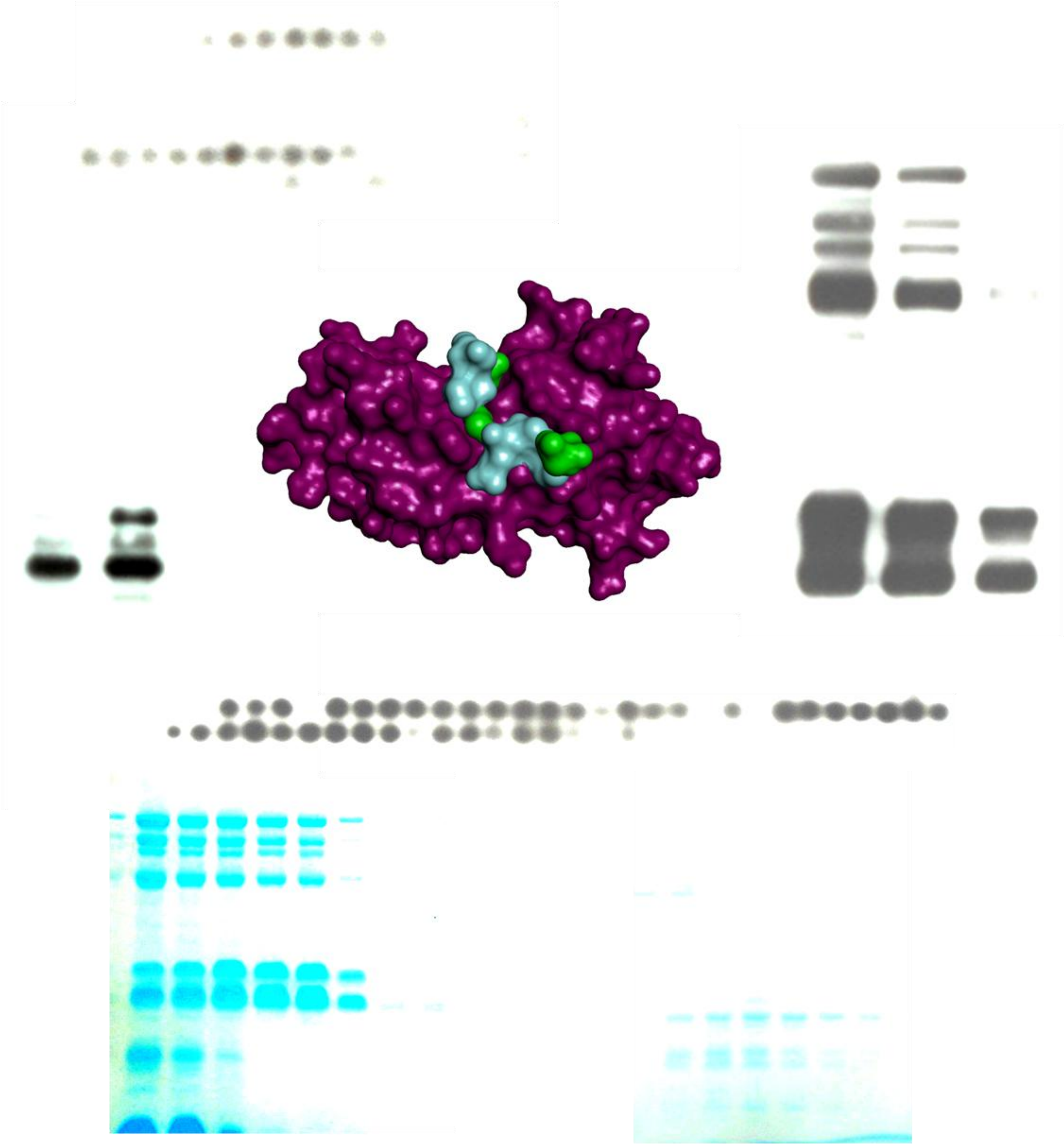
In all predicted EBNA1 dimers, the proline loop of one EBNA1 monomer slots into the central hole, formed between the C-terminal tail and DNA binding core domain of another monomer. Moreover, energy values of the full length dimers are more favourable compared to a dimer constructed using EBNA1 with a deleted C-terminal tail. This suggests a possible role of the proline loop and/or C-terminal tail in the stabilisation of the dimer. In view of these observations, a potential disruptor peptide was designed, the predicted model of which shows a similar topology to the counterpart region (proline loop) of full length EBNA1. However, it lacks a number of potentially critical hydrogen bonds compared to the native proline loop present in full length EBNA1 model. This relaxation in the structure is perhaps due to the absence of spatial constraints, offered by neighbouring residues in the full length protein.

Other studies have successfully shown inhibition of EBNA3C interaction with different components of the cell cycle regulatory complex using a disruptor peptide (Knight *et al*., 2006). Similarly EBNA2 interaction with DNA binding protein CBF1 has also been shown to be inhibited using a peptide derived from the interacting region of CBF1 (Farrell *et al*., 2004). Previously, virtual screening of the 90,000 low molecular weight potential inhibitors (non peptide) for EBNA1-DNA interaction have been reported, demonstrating 3/90,000 of these compounds render reduction in the EBV genome copy numbers and inhibit transactivation function of EBNA1 in a Burkitt's lymphoma cell line (Li *et al*., 2010). In the present study we have proposed a peptide to inhibit its homodimerisation and associated functions. This is the first time a disruptor peptide has been reported for the EBNA1 dimerisation. In molecular docking (blind) DIP interacts within or near the target regions (central hole and curling C-terminal tail). This reflects that the predicted peptide could possibly interfere with the interaction between two EBNA1 monomers by competing with the proline loop for the same site or region.

Short stretch of basic amino acids, also termed protein transduction domain (PTD) have been identified as a carrier of linked cargo such as peptide, protein and nucleic acid into the cell. The TAT sequence, which is derived from the HIV TAT protein, is one such example of PTD (Futaki, 2002; Leifert *et al*., 2003). Treatment of different EBV positive and EBV negative cell lines with the DIP showed efficient intracellular and intranuclear penetration of the peptide. This reflects the efficiency of the TAT sequence as a carrier sequence for cellular transport and also demonstrating that the peptide could reach to the target area (in this case nucleus).

Namalwa cell line harbours 2 integrated copies of EBV genome (Ryan *et al*., 2004) and to date no EBV negative variant has been known for this cell line. This may suggest that EBV genome is an integral part of Namalwa cell line and cell line may not survive or proliferate if the EBV genome has not been maintained properly. As mentioned, genome maintenance of EBV in the infected cells is regulated by EBNA1 dimer, therefore by impairing or destabilizing the dimer formation of EBNA1 by DIP may halt the Namalwa cell proliferation which in turn may lead to their death by apoptosis. Consistently, treatment of DIP showed decline in the Namalwa cells viability compared to untreated and EBV negative BL2 cells. This suggests that DIP may have some intracellular effect on the viability of EBV positive cells at least in Namalwa. Which in light of its origin and molecular docking analyses may be due to its interaction with EBNA1. However, it is important to note that the effect of DIP is observed merely on the trypan blue exclusion assay which may suffer from observational bias and the cell line selected to examine the effect of DIP has least expression of EBNA1 protein. Therefore it is important to replicate the experiment by staining the cells using annexin V/ propium iodide and counting them using flow cytometry. Similarly, direct protein interaction assays could be conducted by immunoprecipitation EBNA1 or DIP from DIP treated Namalwa cell line. Using other cell lines with relatively higher expression of EBNA1 (such as Raji) may also facilitate to optimize the dose required for the peptide to show any potential effect. Moreover, using cell lines with the negative variants of EBV such as BL2, BL30 or BL70, may also tells how efficiently DIP (if it is) can maintain the EBV genome over the period of treatment.

Chapter 4. EBNA1 Protein Protein Interactions



4. Results: EBNA1 Protein Protein Interactions

4.1. Introduction

Several functions of EBNA1 such as genome maintenance, conferring resistance to apoptosis and transactivation are attributed to its ability to interact with different host proteins. In this chapter, the interaction between EBNA1 and two of its partner proteins, EBP2 and USP7, were explored by peptide library array. Human genes encoding EBP2 and USP7 were cloned into bacteria using bacterial protein expression vectors. The recombinant proteins were used as probes over a 25-mer EBNA1 array to explore their potential binding sites on EBNA1. To further resolve the binding regions, the proteins were used as probes over alanine substitution arrays of the binding site regions. Additionally, residues which have been proposed to be phosphorylated were modified accordingly on the arrays to unravel the effect of phosphorylation on binding.

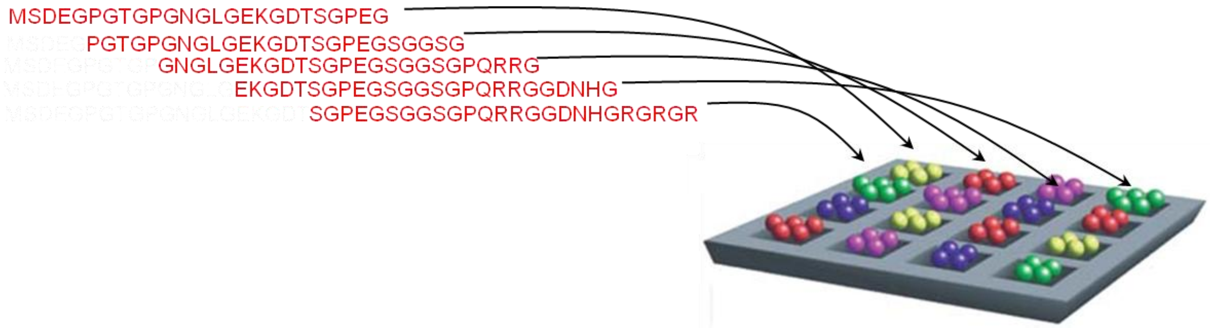
4.2. Validating the system

In order to explore the EBNA1-partner protein interaction, peptide array technology was employed. Peptide arrays of EBNA1 were constructed by synthesizing a library of 25-mer peptides, each shifted along the protein sequence by 5 amino acids (from N-terminal to C-terminal), covering the full length (641 residues) of EBNA1 using either small spots on glass slides or larger spots on cellulose membrane. Peptide synthesis and array spotting were conducted by our collaborators, Ruth MacLeod and Prof. George Baillie. Synthesis and sequential immobilization of the each peptide was conducted following the order of the EBNA1 protein sequence (Figure 4.1A). To verify the synthesis and efficacy of spotting of the peptides on the array, the arrays were probed with three different anti EBNA1 antibodies: Rabbit 16-4, rat IH4 and mouse Aza2E8. The binding of the antibodies on the EBNA1 peptide library was assessed using a western blot development like protocol, where positive binding appears as dark spots on the array. Probing of the EBNA1 array with Rabbit 16-4 showed positive reaction at spots 76-83 which comprise residues: Arg376-Glu435 in EBNA1 sequence. This coincides with the known epitope (personal communication with Prof. J. M. Middeldorp) of this antibody (residues 394-420). Anti EBNA1 IH4 antibody bound to spots 85-88 on the EBNA1 array, which comprises residues Gly421 to Lys460, while Aza2E8 bound to spots 90-92, which correspond to residues Pro446 to Asn480 (Figure 4.2). The epitope of IH4 antibodies has been mapped around 407-450 amino acid on EBNA1 (Snudden *et al.*, 1994) which fits well with its binding on EBNA1 peptide array. The epitope for Aza2E8 epitope is not known, however, due to its ability to impair the EBNA1-DNA interaction, it has been suggested to be within

(A)

MSDEGPGTGPGNGLGEKGDTSGPEGSGGSGPQRRGGDNHGRGRGRGRGRGG
GRPGAPGGSGSGPRHRDGVRRPQKRPSCIGCKGTHGGTGAGAGAGGAGAGGAG
AGGGAGAGGGAGGAGGAGGAGAGGGAGAGGGAGGAGGAGAGGGAGAGGGA
GGAGAGGGAGGAGGAGAGGGAGAGGGAGGAGAGGGAGGAGGAGAGGGAGA
GGAGGAGGAGAGGAGAGGGAGGAGGAGAGGAGAGGAGAGGAGAGGAGGAG
AGGAGGAGAGGAGGAGAGGGAGGAGAGGGAGGAGAGGAGGAGAGGAGGAG
AGGAGGAGAGGGAGAGGAGAGGGGRGRGGSGGRGRGGSGGRGRGGSGGRRGR
GRERARGGSRERARGRGRGRGEKRPRSPSSQSSSSGSPPRRPPPGRRPFFHPVGEAD
YFEYHQEGGPDGEPDVPPGAIEQGPADDPGEGPSTGPRGQGDGGRRKKGGWFGK
HRGQGGSNPKFENIAEGLRALLARSHVERTTDEGTWVAGVFVYGGSKTSLYNLR
RGTALAIPQCRLTPLSRLPFGMAPGPGPQPGPLRESIVCYFMVFLQTHIFAEVLKDA
IKDLVMTKPAPTCNIRVTVCSFDDGVDLPPWFPPMVEGAAAEGDDGDDGDEGGD
GDEGEEGQE

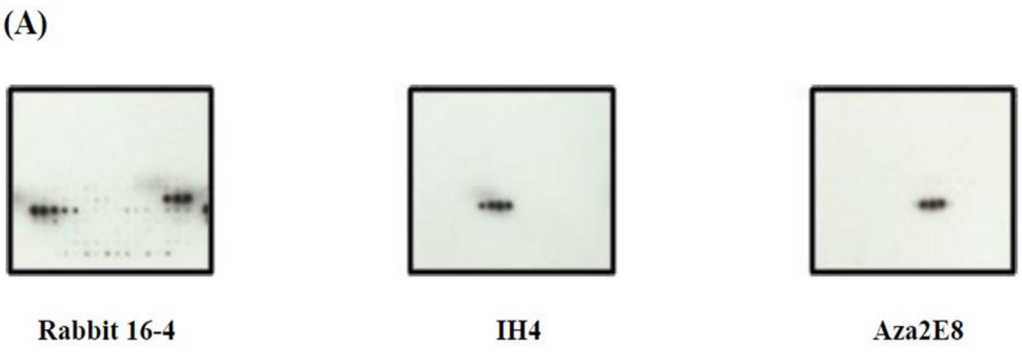
(B)


MSDEGPGTGPGNGLGEKGDTSGPEG
PGTGPGNGLGEKGDTSGPEGSGGSG
GNGLGEKGDTSGPEGSGGSGPQRRG
EKGDTSGPEGSGGSGPQRRGGDNHG
SGPEGSGGSGPQRRGGDNHGRGRGR

(C)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	
	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	
	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	
	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	
	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	
	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	
	113	114	115	116	117	118	119	120	121	122	123	124	125				

Figure 4.1. Peptide Array. The primary protein sequence of EBNA1 (EBV B95-8) is shown in **(A)** where the residues in red, highlights the sequence in the first four spots on the array as indicated below **(B)**. Each spot is composed of 25 amino acids with a 5 amino acids shift from N-terminal to C-terminal. The full map of EBNA1 array covering the full 641 residues of the sequence in 125 spots of peptides is tabulated and shown in **(C)**. The peptide sequences used are given in Appendix II.



(B)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	
	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	
	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	
	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	
	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	
	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	
	113	114	115	116	117	118	119	120	121	122	123	124	125				

(C)

MSDEGPGTGPGNGLGEKGDTSGPEGSGGSGPQRRGGDNHGRGRGRGRGRGGGRPGAPGGSGSGPR
HRDGVRRPQKRPSCIGCKGTHGGTGAGAGAGGAGAGGAGAGGGAGAGGGAGGAGGAGGAGAG
GGAGAGGGAGGAGGAGAGGGAGAGGGAGGAGAGGGAGGAGGAGAGGGAGAGGGAGGAGAG
GGAGGAGGAGAGGGAGAGGAGGAGGAGAGGAGAGGGAGGAGGAGAGGAGAGGAGAGGAGA
GGAGGAGAGGAGGAGAGGAGGAGAGGGAGGAGAGGGAGGAGAGGAGGAGAGGAGGAGAGG
AGGAGAGGGAGAGGAGAGGGGRGRGGSGGRGRGGSGGRGRGGSGGRRGRGRERARGGSRERAR
GRGRGRGEKRPRSPSSQSSSSGSPPRRPPPGRRPFFHPVGEADYFEYHQEGGPDGEPDVPPGAIEQGP
ADDPGEGPSTGPRGQGDGGRRKKGGWFGKHRGQGGSNPKFENIAEGLRALLARSHVERTTDEGTW
VAGVFVYGGSKTSLYNLRRGTALAIPQCRLTPLSRLPFGMAPGPGPQPGPLRESIVCYFMVFLQTHIF
AEVLKDAIKDLVMTKPAPTCNIRVTVCSFDDGVDLPPWFPPMVEGAAAEGDDGDDGDEGGDGDEG
EEGQE

Figure 4.2. Validation of Peptide Array. (A) Peptide arrays were probed with different anti-EBNA1 antibodies as indicated, dark spots on arrays shows the binding of respective antibodies. **(B)** Tabulated representation of the peptide array of EBNA1 is shown, indicating the peptides bound by each antibody: rabbit 16-4 (light green), IH4 (pink) and Aza2E8 (cyan). **(C)** The primary sequence of EBNA1 is shown where residues corresponding to the positive spots are highlighted as Rabbit 16-4 (light green), IH4 (pink) and Aza2E8 (cyan). Light brown shows the overlapping peptide region of Rabbit 16-4 & IH4 binding.

or near to the EBNA1 DNA binding domain (Orlowski *et al*., 1990; Snudden *et al*., 1994) consistent with the observation with these arrays. In summary the binding pattern on the peptide array of the anti EBNA1 antibodies supports the reliability of the spotted peptides for their sequence and order which in turn shows the applicability of the arrays for EBNA1-partner protein interaction studies.

4.3. EBNA1-EBP2 interaction

4.3.1. Expression of human EBP2 in bacterial GST tag expression vector

To investigate the EBNA1-EBP2 binding sites using EBNA1 peptide array, relatively pure EBP2 protein is required. Therefore, a cDNA of human EBP2 was amplified from the Raji cell line and cloned into the bacterial expression vector (pGEX-6P-1), incorporating a glutathione S-transferase tag at N-terminal of the recombinant protein (Appendix III). Clones were verified by colony PCR, restriction digest and DNA sequencing. The strain containing the EBP2 clone vector (denoted herein pEBP2gex) and empty vector (pGEX-6P-1) were grown and induced with IPTG for the expression of the EBP2-GST fusion and GST proteins respectively. Bacterial growth was harvested after overnight incubation and pellets were lysed. Control (pGEX-6P-1) and test (pEBP2gex) bacterial lysates were then purified by affinity chromatography using glutathione sepharose beads. In total six fractions of 500µl were collected and stored at -80^{o}C for subsequent use. Protein concentration of the fractions was measured using BIORAD assay. Aliquots of the fractions were assayed by SDS PAGE and bands were visualized using coomassie blue staining and western blotting using anti-GST antibodies (Figure 4.3).

The comassie blue staining of the fractions of pEBP2gex, showed multiple bands including a band at around 61kDa, expected size of fusion protein (GST-EBP2). In addition, bands at approximately at around 35kDa and 45kDa were observed in the pEBP2gex samples. These may reflect that the fusion protein may have been degraded or translated with multiple truncations yielding the bands of small molecular weights (Figure 4.3). In the pGEX-6P-1 sample the predominant band was appeared at 26kDa, the expected size of GST protein. Both comassie blue stained gels and western blots reflect much lower level of expression of EBP2-GST fusion protein in comparison to alone GST protein (Figure 4.3). Nevertheless, the difference between the protein profile of pEBP2gex and pGEX-6P-1, suggests that fraction 3 and 4 of pEBP2gex may contain adequate amount of the recombinant protein that could be used for subsequent studies.

4.3.2. Probing of EBP2-GST fusion protein on EBNA1 array

In order to explore the binding site of EBP2 on EBNA1, EBNA1 peptide arrays were probed with ~10µg/ml of the EBP2-GST fusion protein (from fraction 4). As control,

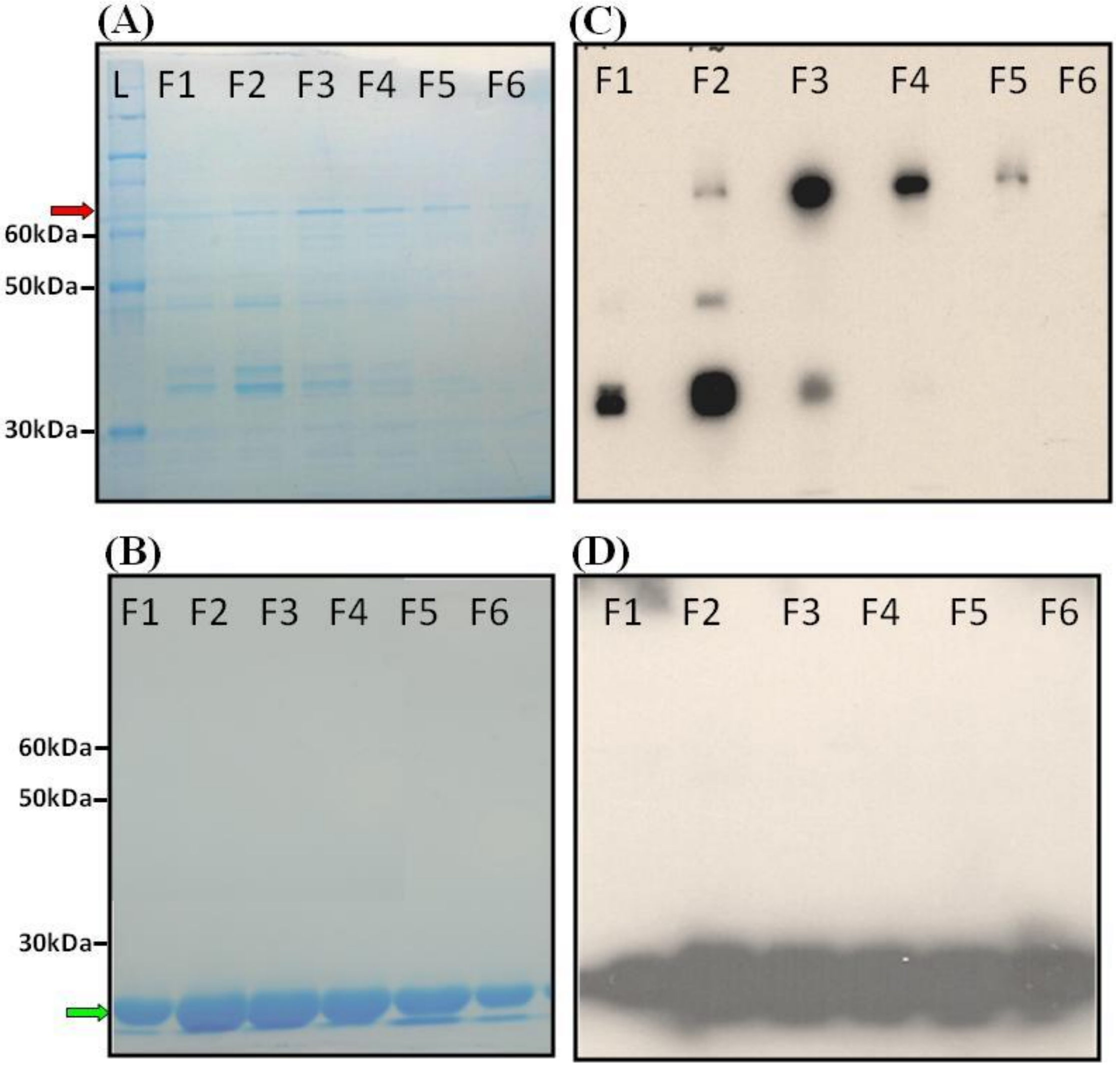


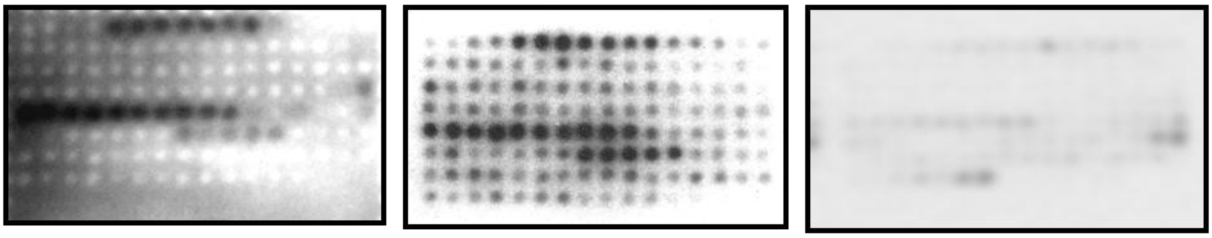
Figure 4.3. Expression of recombinant human EBP2-GST in *E.coli* BL2. Recombinant protein (EBP2-GST) expression was assessed by SDS PAGE. **(A)**: 70µl of each of six 500 µl fractions (F1-F6) of affinity purified EBP2-GST were separated by 10% SDS-PAGE. Proteins were stained with comassie blue. Compared to **(B)** in which same volume of fractions from pGEX-6P-1 were assayed by SDS-PAGE and comassie staining, the pEBP2gex showed bands of the fusion protein (red arrow) whereas fractions from pGEX-6P-1 only showed bands with the expected size of GST (green arrow). **(C)** 20µl of each of six 500 µl fractions (F1-F6) of affinity purified EBP2-GST were separated by 10% SDS-PAGE and western blotted. The blot was probed with antibodies to GST. Compared to **(D)** in which same volume of fractions from pGEX-6P-1 were assayed by SDS-PAGE and western blotted, the pEBP2gex showed bands of the fusion protein whereas fractions from pGEX-6P-1 only showed bands with the expected size of GST.

another EBNA1 array was probed with same amount (~10µg/ml) of GST protein of fraction 2. Both arrays were then developed using anti GST and/or anti EBP2 antibodies (Figure 4.4). On the EBNA1 array, the EBP2-GST fusion protein binding was detected in three strings of spots: spots 5-11, spots 64-74 and spots 88-92, corresponding to the residues Ser21-Lys75, Gly316-Gly390 and Gln436-Asn480 respectively. Two of these bindings sites, Ser21-Lys70 and Gly321-Gly390, fit well with the previously reported EBP2 binding regions on EBNA1 (Shire *et al*., 1999; Nayyar *et al*., 2009). Highlighted in figure 4.4 are the strings of amino acids to which these peptide spots map, in the EBNA1 sequence. Adjacent to the peptide which show strong binding are peptide spots which showed no or weak binding. Excluding the amino acids of these non-binding peptides (weak binding) provide a core binding region, referred to here as the "basic binding region" BBR (Figure 4.4). The BBR of each of the three sites are runs of glycine and arginine residues. Interestingly, the third site which shows relatively weak binding has not been found previously. However, the BBR sequence is GGRRK, again indicating the propensity of this protein to interact with the strings of glycine and arginine. These results appeared to be consistent despite the array being probed with anti GST or anti EBP2 antibodies (Figure 4.4), suggesting the positive spots may indeed reflect the novel binding region for the EBP2 on the EBNA1 molecule. However, probing of the EBNA1 array with GST protein alone showed weak binding to the peptide spots that overlap with those detected by EBP2-GST. Although compared to EBP2-GST fusion protein binding, the binding of GST to the EBNA1 peptide was very weak, it nevertheless raises questions over the reliability EBP2-GST results. To resolve this ambiguity it was decided to switch the vector system from GST-tag to His-tag for the partner protein expression in the subsequent studies.

4.3.3. Expression of human EBP2 in bacterial His tag expression vector

The human EBP2 cDNA was amplified from the pEBP2pgex vector by PCR using primers that incorporated *Hind*III and *Xho*I restriction sites in the DNA. The cDNA was then ligated into the pET-28c vector and transformed into both *E.coli* BL21(DE3) and *E.coli* Rosetta (DE3) strains which incorporate 6xHis tag at the N-terminal of the recombinant gene (Appendix III). The clones were verified by colony PCR, restriction digest and sequencing. Bacteria hosting pEBP2pET28 and pET-28c plasmids were grown and induced with IPTG and after overnight incubation, the bacterial pellets were harvested and lysed. The bacterial lysates were then purified by affinity chromatography using Ni beads. The fractions (5) of 500µl were collected and separated by SDS-PAGE to examine the quantity and quality of the protein by comassie blue staining and western blot (Figure 4.5).

(A)



EBP2-GST (anti EBP2 antibody)

EBP2-GST (anti GST antibody)

GST (anti GST antibody)

(B)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112
113	114	115	116	117	118	119	120	121	122	123	124	125			

(C)

MSDEGPGTGPGNGLGEKGDTSGPEGSGGSGPQRRGGDNHGRGRGRGRGRGGGRPGAPGGSGSGPR
HRDGVRRPQKRPSCIGCKGTHGGTGAGAGAGGAGAGGAGAGGGAGAGGGAGGAGGAGGAGAG
GGAGAGGGAGGAGGAGAGGGAGAGGGAGGAGAGGGAGGAGGAGAGGGAGAGGGAGGAGAG
GGAGGAGGAGAGGGAGAGGAGGAGGAGAGGAGAGGGAGGAGGAGAGGAGAGGAGAGGAGA
GGAGGAGAGGAGGAGAGGAGGAGAGGGAGGAGAGGGAGGAGAGGAGGAGAGGAGGAGAGG
AGGAGAGGGAGAGGAGAGGGGRGRGGSGGRGRGGSGGRGRGGSGGRRGRGRERARGGSRERAR
GRGRGRGEKRPRSPSSQSSSSGSPPRRPPPGRRPFFHPVGEADYFEYHQEGGPDGEPDVPPGAIEQGP
ADDPGEGPSTGPRGQGDGGRRKKGGWFGKHRGQGGSNPKFENIAEGLRALLARSHVERTTDEGTW
VAGVFVYGGSKTSLYNLRRGTALAIPQCRLTPLSRLPFGMAPGPGPQPGPLRESIVCYFMVFLQTHIF
AEVLKDAIKDLVMTKPAPTCNIRVTVCSFDDGVDLPPWFPPMVEGAAAEGDDGDDGDEGGDGDEG
EEGQE

Figure 4.4. EBNA1-EBP2-GST interaction. (A) EBNA1 peptide arrays were probed with 10µg (F3) of human EBP2-GST fusion protein and detected with anti-EBP2 or anti GST antibodies as indicated. Another array was probed with 10µg (F3) of GST protein and probed with anti-GST antibodies as indicated. The dark spots on the arrays reflect positive binding of the proteins. Note the difference in intensity of binding between EBP2-GST fusion protein and GST protein alone. A tabulated map of the peptide array is shown in **(B)**, where cyan coloured boxes correspond to the basic binding region of the fusion protein. The primary sequence of EBNA1 is shown in **(C)** where residues corresponding to the BBR for EBP2 interactions are highlighted as cyan.

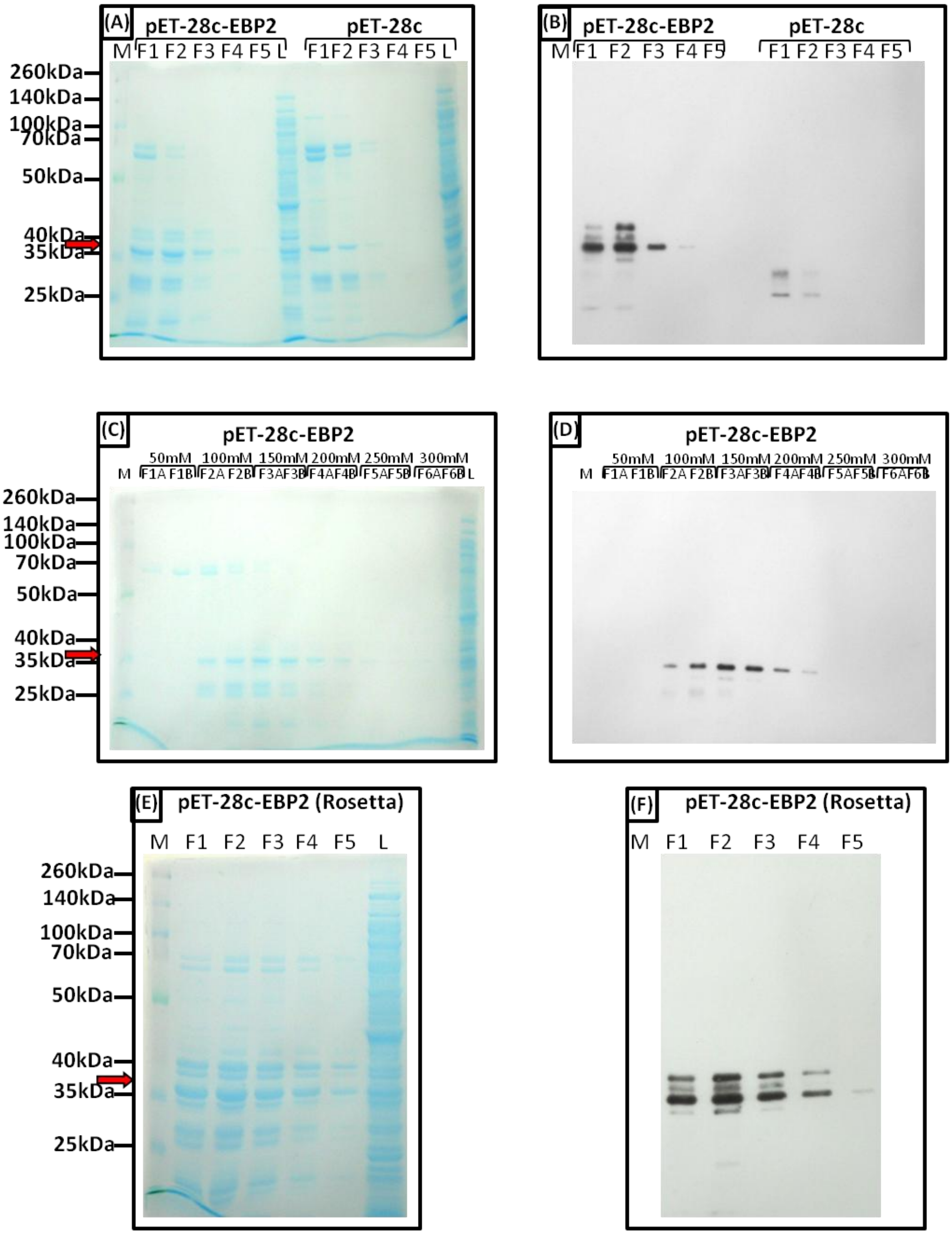


Figure 4.5. Expression of human EBP2 in *E.coli* BL2 and Rosetta strains. Recombinant protein (EBP2-6xHis) were assessed by coomassie blue gel staining (**A**) and western blot (**B**) of the 50µl of fractions (F1-F5), purified using 250mM imidazole from bacterial lysate of pEBP2pET28 and pET-28c (as indicated). Note the difference in the western blot profile of bacteria with EBP2 insert and bacteria with empty vector **(B)**. Fractions eluted by varying the concentration of imidazole in the elution buffer (as indicated) were also examined using coomassie staining (**C**) and western blot (**D**). Fractions purified from lysate of *E.coli* Rosetta(DE3) strain containing pEBP2pET28 bacterial lysate were also examined by comassie blue staining (**E**) and western blots (**F**). 50µl and 10µl from each fraction were used for coomassie gel staining and western blots respectively. For western blots the membranes were probed with anti 6xHis HRP conjugated antibodies. The expected molecular weight (36kDa) of the recombinant protein (EBP2) is indicated by red arrows. Key: M (molecular markers); F (fractions) and L (bacterial lysate).

Protein concentrations of the fractions were estimated by UV absorbance at 280nm. The expected molecular weight of recombinant EBP2-6xHis fusion protein is 36kDa (35kDa of EBP2+ 1kDa of 6xHis). Coomassie blue staining of the gel showed multiple bands of proteins in the different fractions from pEBP2pET28, including prominent bands between 50-70kDa, at ~40kDa, at ~35kDa and at <35kDa. Except the two bands at ~40kDa and a single band at ~35kDa, corresponding fractions from pET-28c clone showed a similar profile on comassie blue staining (Figure 4.5A). The western blot (probed with anti-6xHis antibodies) of a parallel gel clearly distinguished the fractions of both clones (pEBP2pET28 and pET-28c), where prominent bands were observed in fractions from only bacteria containing EBP2 insert at ~35kDa (expected molecular size of EBP2) and between 35-40kDa (Figure 4.5B), while no protein of same molecular weight was detected in fractions from pET-28c clone. This suggests that the ~35kDa protein observed in the bacteria with pEBP2pET28 plasmid was different from the protein of same molecular weight (as observed in the comassie blue stained gel of the respective fractions) in bacteria harbouring empty vector. Similarly, bands (reactive in the western blots; Figure 4.5B), at ~40kDa was only observed in bacteria with recombinant plasmid (pEBP2pET28). These bands may possibly be of full length intact EBP2-6xHis and may have migrated differently on gel electrophoresis whereas <40kDa bands may reflect degradation products of the recombinant protein. Differential migration on the gel could be due to the post translational modification (glycosylation) of the recombinant protein in the bacterial cell. The data show that complete purification was not achieved by affinity chromatography using a single concentration of imidazole (250mM) in the elution buffer. Therefore, the recombinant protein was purified from the bacterial lysate by using increasing concentrations of imidazole in the elution buffer (Figure 4.5C, 4.5D), this resulted in the separation of unwanted bands of 50-70kDa in early fractions (50-150mM imidazole), while bulk of recombinant EBP2-His was eluted in elution buffer containing 100-200mM of imidazole. Western blot of these fractions showed only one band at ~35kDa (expected molecular weight of EBP2) in fractions eluted with 100-200mM imidazole. Both the coomassie blue stained gel and the western blot showed a further degree of purification was achieved using imidazole gradient purification for recombinant EBP2 protein. In order to examine that the less than expected molecular weight (<35kDa) band is due to the interrupted translation of the recombinant protein (e.g. due to the codon bias), plasmid (pEBP2pET28) was transformed in the Rosetta (DE3) strain of *E.coli*. The protein was purified from the clone lysate by affinity chromatography as above. The same EBP2-His bands of 35kDa and 40kDa were produced in the Rosetta strain. This suggests that these proteins were not the result from the interrupted translation due to codon bias (Figure 4.5E, 4.5F). Given that the 6xHis tag is present at the N-terminus of the recombinant protein and these bands were not detected by western of fractions from pET-28c, it is reasonable to suggest that protein of <35kDa are indeed bacterial contaminant proteins.

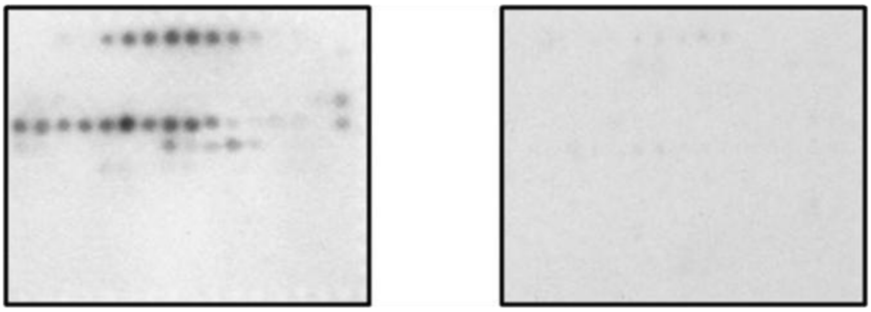
In total, from the comassie blue stained gel and western blot of the fractions, it is clear that the human EBP2 was successfully expressed in the bacteria using his tag expression vector. However, full purification of the recombinant protein was not achieved. Nevertheless the differences between the western blot profile of the fractions from pEBP2pET28 and pET-28c lysates suggest that some fractions of pEBP2pET28 clone (fraction 2/3 from non gradient purification; fractions 3A and 3B from gradient purification) are sufficiently pure by the western detection approach (which is used on the peptide array), therefore could be used for the subsequent EBNA1-EBP2 protein interaction studies.

4.3.4. Probing of EBP2-His protein on the EBNA1 array

EBNA1 glass arrays (analytical arrays) were probed with 7.5µg (30µl) of partially purified recombinant EBP2 protein obtained from fraction no. 2/3 (non gradient). As control, EBNA1 arrays were also probed with 7.5µg protein(s) from fraction 2 from pET-28c clone lysate. Both arrays were developed using anti-6xHis antibodies (Figure 4.6). Probing of recombinant EBP2-His protein on EBNA1 array showed binding at spots: 5-11; 64-74 and 88-91, corresponding the residues: Ser21-Lys75; Gly316-Ser390 and Gln436-Leu485 respectively. The first two binding regions matches well with the reported binding site for EBP2 interaction on EBNA1 (Shire *et al*., 1999; Nayyar *et al*., 2009) and the observations obtained by probing EBNA1 array with the EBP2-GST fusion protein (Figure 4.4). These binding sites resides within the two Gly-Arg rich regions of the EBNA1 protein termed GR-1 and GR-2 (Frappier and O'Donnel1, 1991; Mackey *et al*., 1995). Also consistent with the fusion protein (EBP2-GST) probing, EBP2-His also showed binding to the same spots that correspond to the region previously unknown for EBP2 binding, Gln436-Leu485, indicative of the presence of a potentially novel EBP2 binding region on EBNA1. The control (bacterial lysate) showed no binding on the EBNA1 arrays demonstrating that there is no cross reactivity of either the 6xHis tag or the anti-6xHis antibody with the EBNA1 peptide on the array (Figure 4.6). In addition, EBNA1 array probing with recombinant EBP2 protein from either fraction 2 (non gradient) or fraction 3 (gradient) produced identical results.

In order to explore the EBP2 binding site, an alanine scan array was conducted using a membrane array (preparative array) (Figure 4.7). Peptides which had shown strong and consistent binding by EBP2 fusion protein were selected for this purpose. Briefly, peptide

(A)



(pEBP2pET28) (pET-28c)

(B)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	
	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	
	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	
	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	
	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	
	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	
	113	114	115	116	117	118	119	120	121	122	123	124	125				

(C)

```
MSDEGPGTGPGNGLGEKGDTSGPEGSGGSGPQRRGGDNHGRGRGRGRGRGGGRPGAPGGSGSGPR
HRDGVRRPQKRPSCIGCKGTHGGTGAGAGAGGAGAGGAGAGGGAGAGGGAGGAGGAGGAGAG
GGAGAGGGAGGAGGAGAGGGAGAGGGAGGAGAGGGAGGAGGAGAGGGAGAGGGAGGAGAG
GGAGGAGGAGAGGGAGAGGAGGAGGAGAGGAGAGGGAGGAGGAGAGGAGAGGAGAGGAGA
GGAGGAGAGGAGGAGAGGAGGAGAGGGAGGAGAGGGAGGAGAGGAGGAGAGGAGGAGAGG
AGGAGAGGGAGAGGAGAGGGGRGRGGSGGRGRGGSGGRGRGGSGGRRGRGRERARGGSRERAR
GRGRGRGEKRPRSPSSQSSSSGSPPRRPPPGRRPFFHPVGEADYFEYHQEGGPDGEPDVPPGAIEQGP
ADDPGEGPSTGPRGQGDGGRRKKGGWFGKHRGQGGSNPKFENIAEGLRALLARSHVERTTDEGTW
VAGVFVYGGSKTSLYNLRRGTALAIPQCRLTPLSRLPFGMAPGPGPQPGPLRESIVCYFMVFLQTHIF
AEVLKDAIKDLVMTKPAPTCNIRVTVCSFDDGVDLPPWFPPMVEGAAAEGDDGDDGDEGGDGDEG
EEGQE
```

Figure 4.6. Physical map of EBNA1-EBP2 interaction. (A) EBNA1 peptide arrays were probed with 7.5µg of recombinant human EBP2 protein and bacterial lysate as indicated. The arrays were developed using anti 6xHis antibodies. The dark spots on array indicate the positive binding of the proteins. **(B)** Tabulated representation of the peptide array of EBNA1, where cyan coloured boxes correspond to the binding region of the EBP2-His on the EBNA1 array. **(C)** The primary sequence of EBNA1 is shown where residues corresponding to the positive spots for EBP2 interactions (representing BBR) on array are highlighted in cyan.

(A)

Peptide 8. 36-GDNHGRGRGRGRGRGGGRPGAPGGS-60

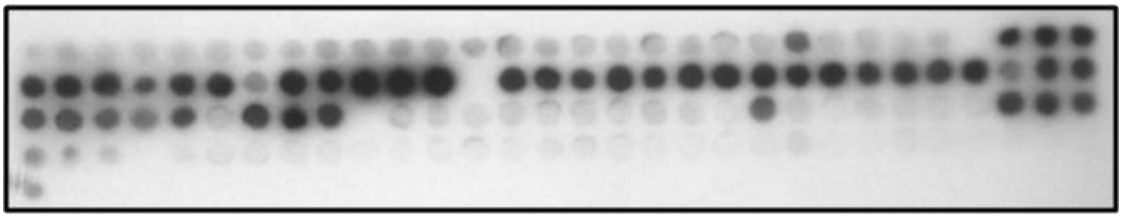
Peptide 10. 46-GRGRGGGRPGAPGGSGSGPRHRDGV-70

Peptide 70. 346-GRGGSGGRRGRGRERARGGSRERAR-370

Peptide 73. 361-ARGGSRERARGRGRGRGEKRPRSPS-385

Peptide 91. 451-RGQGDGGRRKKGGFGKHRGQGGSN-475

(B)



(C)

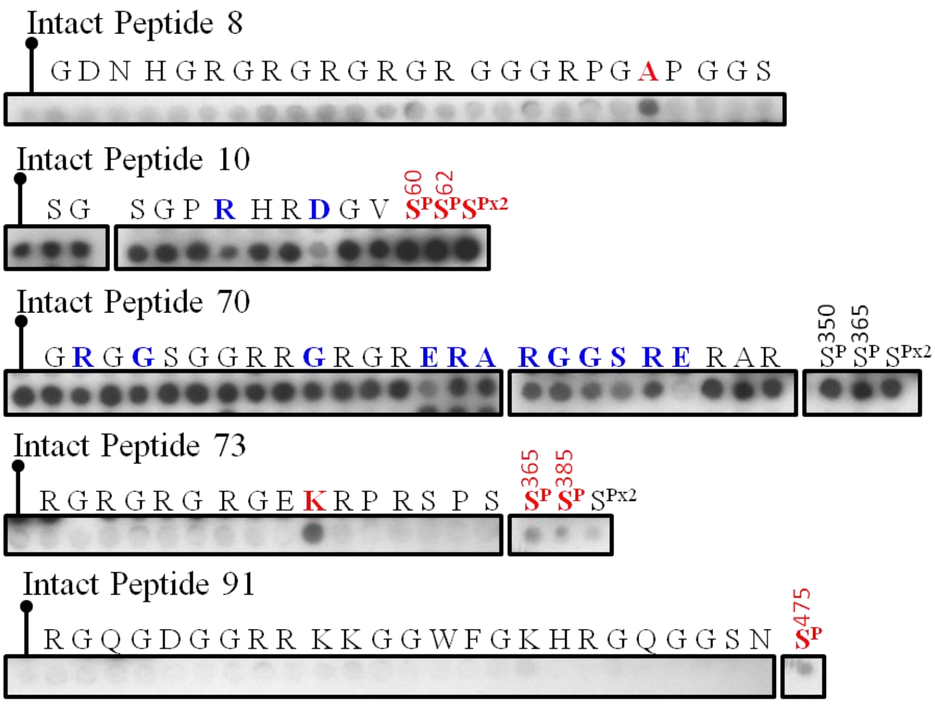

Intact Peptide 8
G D N H G R G R G R G R G R G G G R P G A P G G S
Intact Peptide 10
S G S G P R H R D G V S^P S^P S^Px2
60 62
Intact Peptide 70
G R G G S G G R R G R G R E R A R G G S R E R A R S^P S^P S^Px2
350 365
Intact Peptide 73
R G R G R G R G E K R P R S P S S^P S^P S^Px2
365 385
Intact Peptide 91
R G Q G D G G R R K K G G W F G K H R G Q G G S N S^P
475

Figure 4.7. Exploring critical residues involved in the EBNA1-EBP2 interaction. EBNA1 alanine scanning arrays for the EBP2 binding sites **(A)** were spotted and probed with 45µg (200µl) of recombinant human EBP2-His **(B)**. Each peptide on the membrane is separated and shown in **(C).** The first spot in each row represent the unmodified peptide sequence spot as indicated. The following spots are the derivatives of the wild type sequence with successive amino acids substituted with Ala (or Asp in the case of Ala been present in wild type sequence). Peptides with one or more phosphorylated residue (superscript P) were also assayed. Ala scan of peptide 10 and 73 started from residue 60 and 370 respectively **(C)**. The amino acids in blue and red indicate the decrease or increase in EBNA1-EBP2 binding due to the amino acid substitution respectively.

8 (corresponding residues: Gly36-Ser60), 10 (corresponding residues: Gly46-Val70), peptide 70 (corresponding residues: Gly346-Arg370), peptide 73 (corresponding residues: Ala361-Ser385) and peptide 91 (corresponding residues: Arg451-Asn475) were synthesized and spotted on the membrane array (An MRes student in the laboratory, Jin, contributed to the conducting this set of experiments). In addition to the wild type sequences, variants of each were also spotted with successive single amino acid substitutions to Ala (or Asp in case of Ala being present in the wild type sequence), from the N-terminus to C-terminus of the peptide (Figure 4.7). The array was probed with 45µg (200µl) of the recombinant EBP2 protein. In contrast to glass slide array, the peptide 8 showed very weak binding with EBP2, however, on the alanine scan, Asp56Ala substitution yielded an increased signal for binding. Consistent with the glass array results, spot 10 (residues: Gly46-Val70) which also corresponds to the N-terminal EBP2 binding site (at GR1), showed strong binding. Alanine scan of this site showed reduced binding with the Arg65Ala substitution and the binding was considerably reduced with the Asp68Ala substitution, indicative of the importance of these residues in EBNA1 N-terminal binding of EBP2. Interestingly, significant improvement in the EBP2 binding was observed when either one or both serines in peptide 10 (Ser60 and Ser62) were replaced by their phosphorylated versions, suggesting that phosphorylation of serine may positively affects the binding of EBP2 with the N-terminus of EBNA1, however, it is important to note that both of these serines reside outside the reported N-terminal binding site for EBP2-His on EBNA1 (Shire *et al.*, 1999). Similarly the Arg65 and Asp68 are also not included in the previously reported EBNA1 binding region for EBP2. Peptide 70, which corresponds to main EBP2 binding site (residues: Gly346-Arg370), on probing with EBP2 showed strong binding. Peptide 70 and several of its derivatives showed a similar level of binding by EBP2-His as did peptide 10. However, substitutions at: Arg347Ala, Gly349Ala, Gly355Ala, Arg356Ala, Arg360Ala and Arg366Ala all reduced the EBP2-His interaction while substitutions at Glu359Ala and Glu367Ala nearly abolished EBP2-His interactions. The data suggest these may be the critical residues in the GR2 region of EBNA1 for binding with EBP2. In contrast to GR1 region, phosphorylation of Ser350 and Ser365 did not show any effect on EBP2-His interaction with the peptide. Although peptide 73 and 91 (corresponding to the residues in the potentially novel site for EBP2 binding) showed nearly consistent binding on all EBNA1 glass arrays on probing with EBP2-His/EBP2-GST, on the membrane array spots, the recombinant protein failed to show strong binding to the respective sequence spot and its alanine scan. An exception in this regard is the substitution, Lys379Ala, which noticeably improved the binding of EBP2 on the respective spot. Also incorporation of phosphorylated Ser365 and Ser385 on peptide 73 and Ser474

on peptide 91 enhanced the binding by EBP2 on the array compared to the unmodified peptide.

The present peptide array approach supports the previous findings relating to the physical region on EBNA1 for EBP2 binding determined on the basis of deletion mutation and yeast two hybrid assays (Shire *et al*., 1999; Kappor *et al*., 2001; Nayyar *et al*., 2009). In addition, the data also points to certain key residues and possibly the involvement of post translation modification in EBNA1-EBP2 interactions.

4.4. EBNA1-USP7 interaction

4.4.1. Expression of human USP7 in bacterial His tag expression vector

A human USP7 cDNA clone was obtained from Addgene. The cDNA was transferred to the pET-28c expression vector by PCR, incorporating *Hind*III and *Xho*I restriction sites at 5’ and 3’ ends respectively (Appendix III). This was transformed into *E.coli* BL21 (DE3) and *E.coli* Rosetta (DE3) strains. The clones were verified by colony PCR, restriction digest and sequencing. Bacteria containing pUSP7pET28 and pET-28c plasmids were grown and induced with IPTG at 37^{o}C and after overnight incubation, the bacterial pellet was harvested and lysed. The bacterial lysate was then purified through affinity chromatography using Ni beads. Fractions of 500µl were collected and aliquots were electrophoresed to examine the quantity and quality of the protein by comassie staining and western blot (Figure 4.8). Protein concentrations of the fractions were estimated by UV absorbance at 280nm. The expected molecular weight of the recombinant USP7-His is 129kDa (128kDa of USP7 + 1kDa 6xHis). Comassie blue gel staining of the fractions (from pUSP7pET28 containing bacteria) showed multiple bands, including prominent bands between 100kDa and 140kDa (expected molecular size of USP7), a band at approximately 75kDa, and two prominent bands at ~42kDa, and ~37kDa. The lysate of pET-28c clone was processed similarly and the comassie blue stained gel profile of the fractions showed a faint band at ~100kDa and bands at approximately ~70kDa, 68kDa, ~35kDa and ~27kDa in fractions 1-3 (Figure 4.8A). Western blot of parallel gels showed reactivity with multiple bands present in the fractions of pUSP7pET28 (similar to those seen by comassie blue staining) insert while only bands <30kDa were detected in the fractions from pET-28c clone. In addition to the bands between 100kDa and 140kDa (expected molecular weight of USP7), low molecular weight bands were also observed in the fractions of lysate from pUSP7pET28 clone, especially between 70-100kDa, ~40kDa and between 40kDa and 35kDa, suggesting the possibility of truncated protein synthesis and/or degradation of recombinant protein (Figure 4.8B). The data show that the complete purification of the recombinant USP7 was not achieved by affinity chromatography

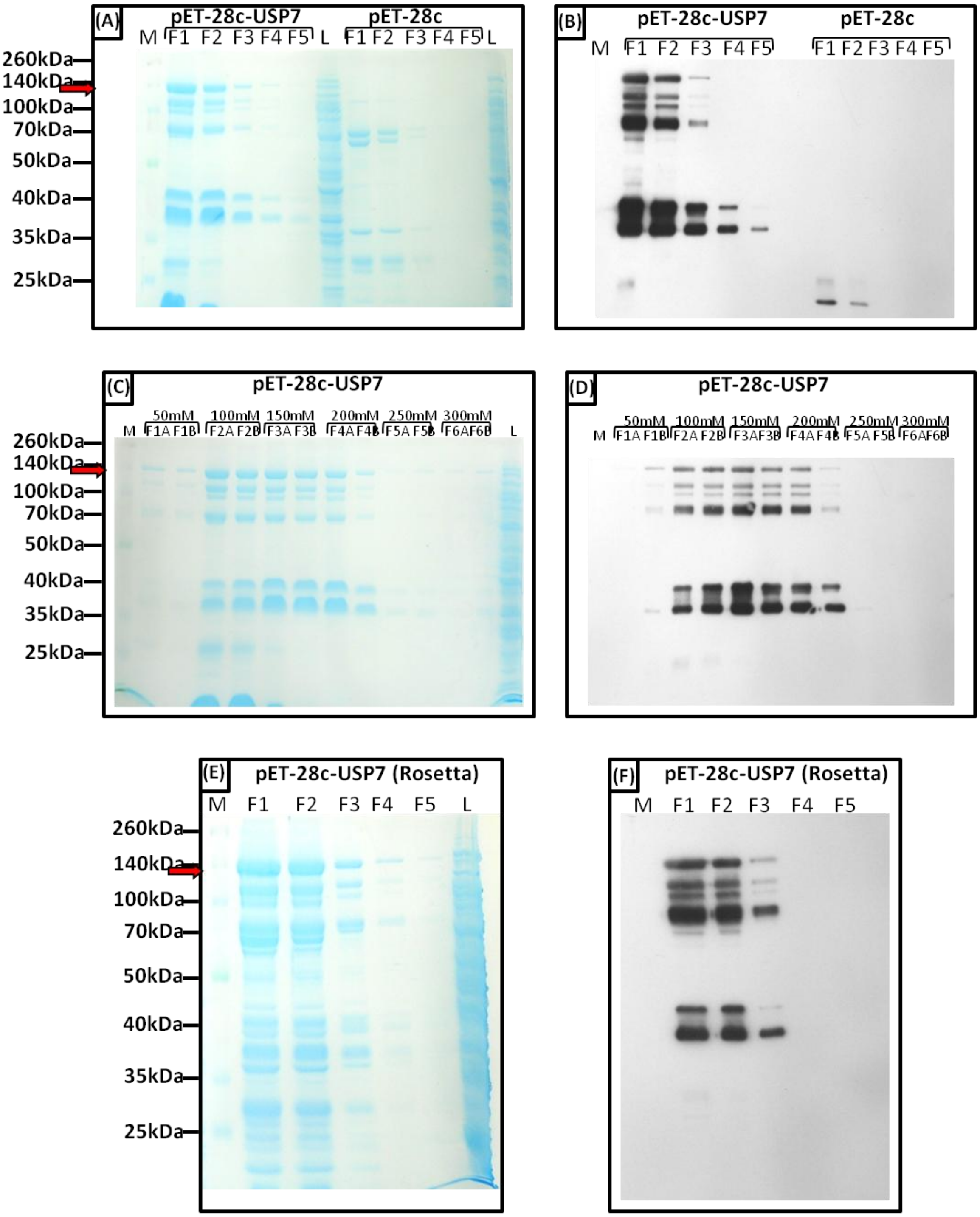

(A)
pET-28c-USP7
pET-28c
M F1 F2 F3 F4 F5 L F1 F2 F3 F4 F5 L
260kDa
140kDa
100kDa
70kDa
50kDa
40kDa
35kDa
25kDa
(B)
pET-28c-USP7
pET-28c
M F1 F2 F3 F4 F5 F1 F2 F3 F4 F5
(C)
pET-28c-USP7
50mM 100mM 150mM 200mM 250mM 300mM
M F1A F1B F2A F2B F3A F3B F4A F4B F5A F5B F6A F6B L
(D)
pET-28c-USP7
50mM 100mM 150mM 200mM 250mM 300mM
M F1A F1B F2A F2B F3A F3B F4A F4B F5A F5B F6A F6B
(E)
pET-28c-USP7 (Rosetta)
M F1 F2 F3 F4 F5 L
(F)
pET-28c-USP7 (Rosetta)
M F1 F2 F3 F4 F5

Figure 4.8. Expression of human USP7 in *E.coli* BL2 and Rosetta strains. Recombinant protein (USP7-His) was assessed by coomassie blue gel staining **(A)** and western blot **(B)** of the fractions (F1-F5) from BL2, purified using 250mM imidazole from lysates of pUSP7pET28 and pET-28c (as indicated). Note the difference in the western blot profile of bacteria with USP7 insert and bacteria with empty vector. Fractions eluted using elution buffer of different concentrations of imidazole (as indicated) were also examined using coomassie gel **(C)** and western blot **(D)**. Fractions purified from the *E.coli* Rosetta strain containing pUSP7pET28 bacterial lysate were also examined by comassie blue staining **(E)** and western blots **(F)**. 50µl and 10µl from each fraction were used for comassie blue gel staining and western blots respectively. For western blots the membranes were probed with anti 6xHis HRP conjugated antibodies. The expected molecular weight (129kDa) of the recombinant protein (USP7) is indicated by red arrows. Key: M (molecular markers); F (fractions) and L (bacterial lysate).

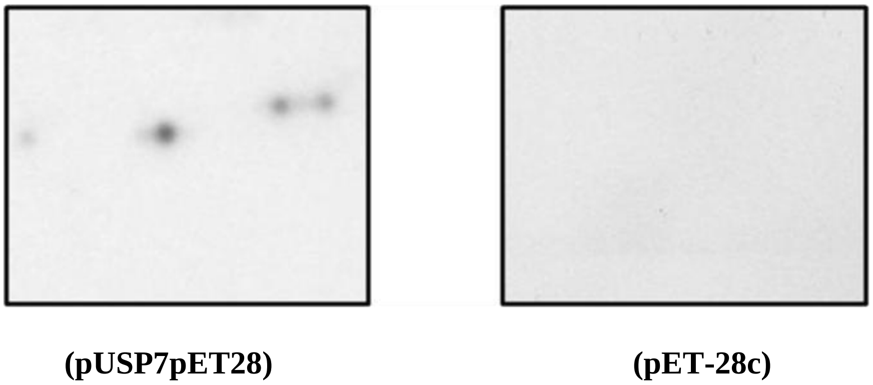
using a single concentration of imidazole (250mM) in the elution buffer. Therefore, the recombinant protein (USP7) was purified again using elution buffers with increasing concentrations of imidazole (Figure 4.8C, 4.5D). Both comassie gel and western blot of the fractions obtained in this regard, showed no significant improvement in the purification of the recombinant USP7 as both expected and unexpected molecular weight protein bands appeared proportionally in all fractions eluted using 50mM to 300mM of imidazole. In order to investigate if the lower than expected molecular weight (<129kDa) bands result from interrupted protein synthesis (due to codon bias) of the recombinant protein, the same plasmid (pUSP7pET28) was transformed into the Rosetta strain (DE3) of *E.coli*. The protein from the clone lysate was purified by affinity chromatography. Both comassie blue stained gel and western blot showed some improvement in procuring the potentially full length recombinant USP7 as the bands between 35-40kDa became less intense without losing the intensity of bands between 100kDa and 140kDa (Figure 4.8E, 4.8F). In order to investigate if these fragments could be readily separated, extracts were subjected to size exclusion chromatography. Comassie blue gel staining (not shown) of the fractions showed that all the fragments appear to elute with the same profile and not according to their size by gel. This might reflect that the fragmented recombinant protein is structurally held together and only become separated at denaturation for separation by SDS-PAGE.

In summary, from the comassie blue stained gel and western blot of the fractions, it is clear that the human USP7-His was successfully expressed in the bacteria. However, complete purification of the recombinant protein was not attained. Nevertheless the difference in the western blot profile of the fractions between the bacteria containing pUSP7pET28 and empty vector suggests some fractions (F2, F2A-F3B) from the bacteria hosting pUSP7pET28 plasmid could be used for the subsequent studies.

4.4.2. Probing of USP7-His protein on the EBNA1 array

EBNA1 glass arrays were probed with 20µg (22µl) of partially purified recombinant USP7 protein from fraction no. 2 (from BL2 strain). As control, EBNA1 arrays were also probed with 20µg (50µl) from fraction 2 of lysate from bacteria containing empty vector. Both arrays were developed using anti-6xHis antibodies (Figure 4.9). Probing of the EBNA1 array with recombinant USP7-His revealed strong binding to spots: 78, 80 and 87 (particularly the latter) corresponding the residues: Ser386-Gly410; Arg396-Glu420 and Pro431-Asp455 respectively. The reported binding site for USP7 interaction on EBNA1 ranges from Gln436 to Pro450 (Holowaty *et al.*, 2003; Saridakis *et al.*, 2005). This sequence can be completely found in peptide 86 to 88. The control fraction showed no

(A)



(B)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	
	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	
	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	
	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	
	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	
	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	
	113	114	115	116	117	118	119	120	121	122	123	124	125				

(C)

MSDEGPGTGPGNGLGEKGDTSGPEGSGGSGPQRRGGDNHGRGRGRGRGRGGGRPGAPGGSGSGPR
HRDGVRRPQKRPSCIGCKGTHGGTGAGAGAGGAGAGGAGAGGGAGAGGGAGGAGGAGGAGAG
GGAGAGGGAGGAGGAGAGGGAGAGGGAGGAGAGGGAGGAGGAGAGGGAGAGGGAGGAGAG
GGAGGAGGAGAGGGAGAGGAGGAGGAGAGGAGAGGGAGGAGGAGAGGAGAGGAGAGGAGA
GGAGGAGAGGAGGAGAGGAGGAGAGGGAGGAGAGGGAGGAGAGGAGGAGAGGAGGAGAGG
AGGAGAGGGAGAGGAGAGGGGRGRGGSGGRGRGGSGGRGRGGSGGRRGRGRERARGGSRERAR
GRGRGRGEKRPRSPSSQSSSSGSPPRRPPPGRRPFFHPVGEADYFEYHQEGGPDGEPDVPPGAIEQGP
ADDPGEGPSTGPRGQGDGGRRKKGGWFGKHRGQGGSNPKFENIAEGLRALLARSHVERTTDEGTW
VAGVFVYGGSKTSLYNLRRGTALAIPQCRLTPLSRLPFGMAPGPGPQPGPLRESIVCYFMVFLQTHIF
AEVLKDAIKDLVMTKPAPTCNIRVTVCSFDDGVDLPPWFPPMVEGAAAEGDDGDDGDEGGDGDEG
EEGQE

Figure 4.9. EBNA1-USP7 interaction. (A) EBNA1 peptide array was probed with 20µg of recombinant human USP7 protein and bacterial lysate as indicated. The arrays were developed using anti 6xHis antibodies. The dark spots on array indicate the positive binding of the protein. **(B)** Tabulated representation of the peptide array of EBNA1 where cyan coloured boxes correspond to the binding spots of the recombinant USP7 protein on the EBNA1 array. **(C)** The primary sequence of EBNA1 is shown where residues corresponding to the positive spots for USP7 interaction on array are highlighted as yellow (peptide 78) and cyan (peptide 87).

(A) Peptide 87. 431-PGAIEQGPADDPGEGPSTGPRGQGD-455

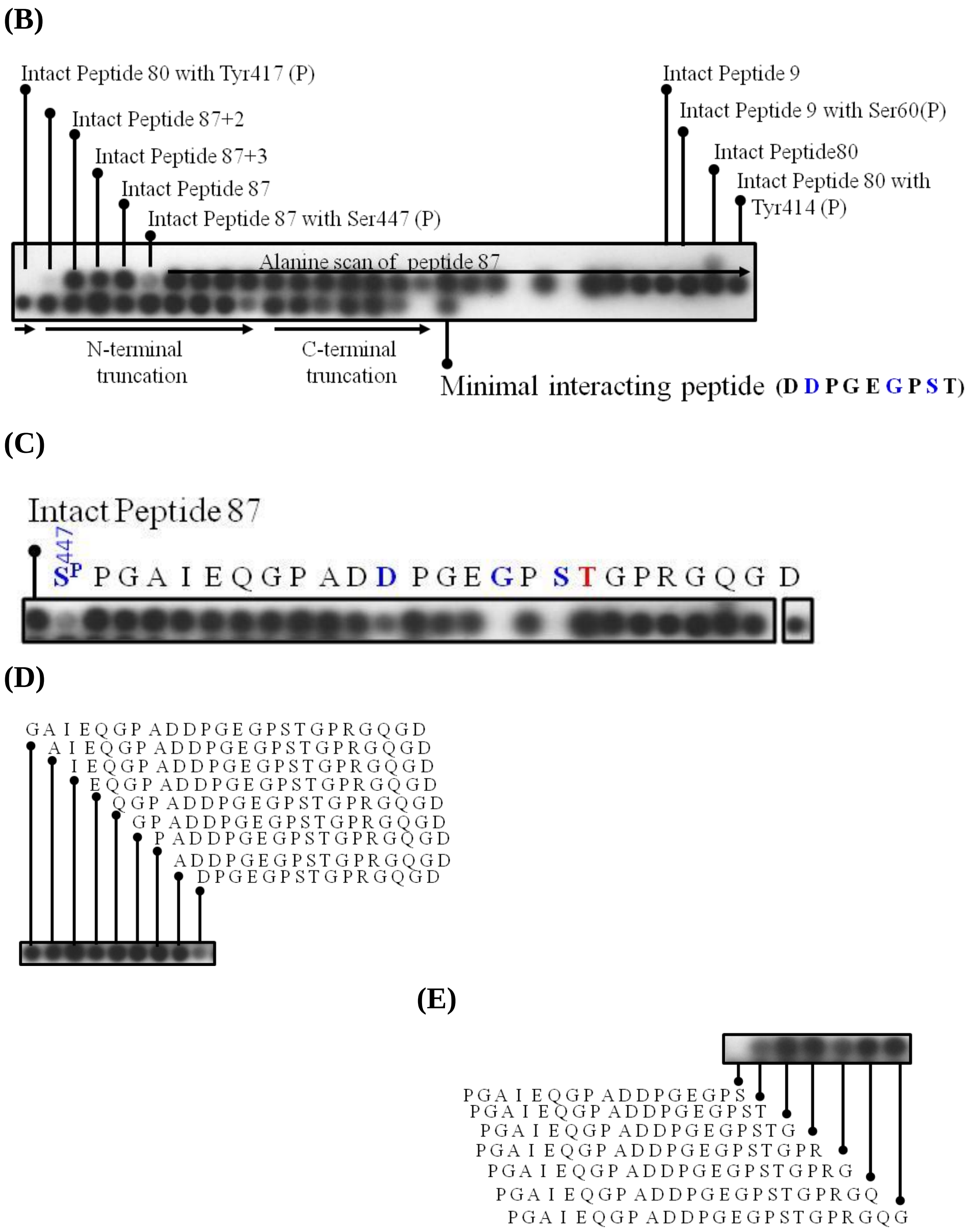

(B)
Intact Peptide 80 with Tyr417 (P)
Intact Peptide 87+2
Intact Peptide 87+3
Intact Peptide 87
Intact Peptide 87 with Ser447 (P)
Intact Peptide 9
Intact Peptide 9 with Ser60(P)
Intact Peptide80
Intact Peptide 80 with Tyr414 (P)
Alanine scan of peptide 87
N-terminal truncation
C-terminal truncation
Minimal interacting peptide (D D P G E G P S T)
(C)
Intact Peptide 87
447
S P P G A I E Q G P A D D P G E G P S T G P R G Q G D
(D)
G A I E Q G P A D D P G E G P S T G P R G Q G D
A I E Q G P A D D P G E G P S T G P R G Q G D
I E Q G P A D D P G E G P S T G P R G Q G D
E Q G P A D D P G E G P S T G P R G Q G D
Q G P A D D P G E G P S T G P R G Q G D
G P A D D P G E G P S T G P R G Q G D
P A D D P G E G P S T G P R G Q G D
A D D P G E G P S T G P R G Q G D
D P G E G P S T G P R G Q G D
(E)
P G A I E Q G P A D D P G E G P S
P G A I E Q G P A D D P G E G P S T
P G A I E Q G P A D D P G E G P S T G
P G A I E Q G P A D D P G E G P S T G P R
P G A I E Q G P A D D P G E G P S T G P R G
P G A I E Q G P A D D P G E G P S T G P R G Q
P G A I E Q G P A D D P G E G P S T G P R G Q G

Figure 4.7. Critical residues involved in EBNA1-USP7 interaction. Alanine scan of the EBNA1 peptide 87 (A) was spotted and was probed with 125µg of recombinant human USP7 protein (B). The first spot in (C) represents the corresponding wild type sequence (peptide 87) spot as indicated. The following spots are the progenies of the peptide 87 where single amino acid is substituted with Ala or Asp (in case of alanine present in the native sequence) in each spot as indicated. A peptide with phosphorylated Ser is also spotted on the array (indicated by superscript P). Probing with the N-terminal and C-terminal truncated peptides are shown in (D) and (E) with the spotted peptide indicated.

binding on the EBNA1 array reflecting the reliability of the approach (Figure 4.9).

In order to further delineate the USP7 binding site on EBNA1, an alanine scan of peptide 87 was generated and tested. For this purpose, wild type sequence (corresponding to the spot 87; Figure 4.10A) was spotted on the membrane array. In addition, to examine the effects of peptide selection, 3 spots, shifting this sequence each by one amino acid (i.e. 87+1, 87+2 and 87+3) were spotted. Also, to compare this to spot 80 binding, peptide 80 with each (in turn) of the tyrosine phosphorylated were included. Additionally, derivatives of the wild type peptide 87 sequence were included where each amino acid in turn was substituted with Ala (Asp in the case where Ala is present in the wild type sequence) from the N-terminal (Pro) to the C-terminal (Asp). To define the minimal region on EBNA1 for interaction with USP7, further derivatives of peptide 87 were synthesized and spotted sequentially truncating from the N-terminus and C-terminus one amino acid at a time. The complete set of the array was probed with 125µg (130µl) of the recombinant USP7-His protein (fraction 2 from BL2) and developed by probing anti His antibodies (Figure 4.10B). On this array peptide 80 showed very weak binding (in comparison to peptide 87) and this was abolished when either of the tyrosines was phosphorylated. Shifting the frame of peptide 87 by one amino acid at a time gave weaker binding for 87+1 but good binding with 87+2 and 87+3, relative to peptide 87. Shifting by 5 amino acids either way (peptide 86 or 88) showed only very weak binding on the original glass slide array. As expected a strong binding was observed for the unmodified peptide 87. The array showed that a substitution at Asp441Ala weakened the recombinant USP7-His binding with the EBNA1 peptide whereas substitutions at Gly445Ala and Ser447Ala abolished the USP7-His interaction with the peptide. Additionally, phosphorylation at Ser447 also substantially decreased the binding of USP7 to the EBNA1 peptide compared to the unmodified peptide (Figure 4.10C). N-terminal truncation of the peptide showed no effect upon USP7-His binding to the peptide from deletion of Pro431 to Ala438, however the binding significantly dropped upon the further deletion of Asp440 (Figure 4.10D). At the C-terminal end, no effect on the interaction was observed through the successive deletion of Asp455 to Gly449 however, the further deletion of Thr448 reduced the binding and deletion of Ser447 completely abolished the peptide 87-USP7-His interaction (Figure 4.10E). The data collectively shows that the minimal region that is important for the EBNA1-USP interaction spans from Asp440-Thr448 on the EBNA1 sequence. Furthermore Asp 441, Gly445 and Ser447 were identified as critical residues in this interaction with USP7. The minimal binding region peptide alone showed good binding, despite the reduced binding seen with peptides 87+1, 86 and 88, which all contain this sequence.

4.5. Summary of findings

- Binding of anti-EBNA1 antibodies to the peptide array spots corresponding to the specific epitopic regions reflects the sequential specificity of EBNA1 on the array.
- Binding of the EBNA1 epitope on the array with the conformation sensitive antibody (Aza2E8) suggests that the peptides on the array have atleast some correct structural attributes.
- EBNA1 partner proteins, EBP2 and USP7, were successfully expressed as his tag fusions using a bacterial expression system, however absolute purification of the recombinant proteins would require an additional battery of protein purification techniques. Nevertheless, both recombinant proteins were detected on western blots suggesting that fractions obtained by purification could be exploited for peptide array studies.
- Binding of recombinant EBP2 to EBNA1 arrays was consistent with the reported binding sites (GR1 and GR2 region).
- A novel binding site of EBP2, downstream (~75 amino acids) to the GR2 region, was observed in the slide array but not in membrane array and this novel site is in close proximity to the GR2 region in the three dimensional model of EBNA1.
- Substitutions of some positively charged residues or negatively charged residues of EBNA1 with hydrophobic residues, especially in the GR2 region increased or decreased the binding of recombinant EBP2 with the EBNA1 peptides (respectively). This suggests that electrostatic repulsion between the arginines of GR2 may negatively regulate the binding of EBP2 with EBNA1.
- Phosphorylation of serine residues, especially in the GR1 region enhanced the binding of recombinant EBP2 to EBNA1 peptides, suggestive of the positive regulation of EBNA1-EBP2 binding by post translation modification.
- Akin to EBP2, USP7 binding with the EBNA1 array was also consistent with the reported binding region. A novel binding site was only observed in the slide arrays.
- An alanine scan of an EBNA1 peptide which interacts with USP7 revealed the minimal binding region required for this interaction and also highlighted critical residues in this intermolecular binding.
- Phosphorylation of serine residues in the USP7 binding site of EBNA1 decreased the binding of recombinant USP7 suggesting that phosphorylation of the serine may negatively regulate the interaction.

4.6. Discussion

Investigating protein protein interactions is central to developing a working understanding of biological systems. Peptide array represent a new methodology to study protein protein interactions *in vitro*. To date the method has been successfully employed in many studies such as studying the PDE4D5-β and arrestin interaction (Baillie *et al*., 2007), PAX protein interactions (Okada *et al*., 2012) and Argonuate-GW182 interactions (Pfaff *et al*., 2013). Despite being highly promiscuous in partner binding, this approach has not been used to explore the interaction of EBNA1 with its partner proteins. In the present study we explored the validity of this system for protein association studies of EBNA1.

Structural conformations of EBNA1 peptides on array

The binding of anti EBNA1 antibodies to their specific epitopes validates the EBNA1 protein sequence on the peptide array (Figure 4.2). Of these, Aza2E8 binding with EBNA1 has been suggested to be conformation sensitive (Hearing *et al*., 1985). Therefore binding of Aza2E8 to a 25-mer peptide on the peptide array suggests the presence of native structural conformation of peptides on the array. We do not know whether this observation could be generalized to the other peptides spotted on the array or limited to the spots where Aza2E8 binds. One way of evaluating would be to use a conformational sensitive set of antibodies that covers the full length of the EBNA1 molecule (but such a set is not available).

Heterologous expression of EBNA1 partner proteins in bacteria

For protein protein interaction studies using peptide array, partially purified protein is preferential for challenging the binding partner array. In this study, two human proteins, EBP2 and USP7 were expressed in *E.coli* BL2 using the bacterial expression vector (pGEX-6P-1 and pET-28c). Though both proteins were expressed successfully in the bacteria, complete purification was not achieved in either case. First in both cases, higher than expected molecular weight proteins were observed. These differences could be explained in terms of difference in the migration rate of the marker proteins and recombinant protein due to the presence of imidazole in the latter, which may resist the flow of current. Alternatively, it is not uncommon for protein to migrate differently from their expected size, due to conformational attributes. It is also possible that these proteins may be bacterial protein contaminants, however the respective size bands were not observed in comassie stained gel and western blot of the fractions of bacterial lysate from bacteria containing empty vector (pET-28c). Additionally, the gels of corresponding fractions were probed with the anti EBP2 and anti USP7 antibodies (data not shown) which supports their identity as EBP2 and USP7 respectively. This problem could be resolved by removing the imidazole from the fractions by dialysis, however, it may result in the significant loss of the protein (possibly variably from each of the fractions). Additionally, mass spectrometric analyses of the proteins of unexpected size can reveal their true identity.

Comassie blue staining and western blots of the fractions obtained from bacteria with EBP2 and USP7 inserts showed multiple small size (in comparison to expected) protein bands. In some cases similar sized bands were also observed in the comassie blue stained gel of the fractions from bacteria with empty vector which indicates that some small bands are due to the contamination with bacterial proteins during purification. However, none of those bands were found to be responsive to anti-His antibodies in western blotting. The small size bands (smaller than expected for recombinant proteins) detected are in Westerns blots are therefore either the degraded products of the recombinant protein or are the result of truncated protein synthesis. Although *E.coli* BL21 lacks many of the bacterial proteases it still has some proteases which could potentially be resistant to the enzyme inhibitors used in the lysis buffer and may degrade the recombinant protein before or after the lysis. As these human proteins (EBP2 and USP7) have been expressed in bacteria which lack the guidance machinery for protein folding (chaperone system), the recombinant proteins may have been misfolded. This misfolding of the recombinant protein may result in thermodynamic instability of the proteins, which in turn may promote their degradation despite the lack/inhibition of major proteolytic enzymes. In addition to lysozyme, sonication was used for the bacterial lysis which may also lead to the degradation of proteins due to the generation of heat. In the present study, both altering the sonication conditions and protease inhibitor composition and concentration did not make any noticeable impact on the purification profile of recombinant proteins (data not shown). Alternatively, it is possible that the expression of the recombinant protein was hampered due to the differences in the codon preferences between prokaryotes and humans. This negative effect on the heterelogous protein expression could be due to the translational frame shifting (Spanjaard and van Duin, 1988), premature termination of translation (Gerchman *et al*., 1994), amino acid misincorporation (Calderone *et al*., 1996) and in frame translational hop (Kane *et al*., 1992). The human EBP2 gene of 306 codons contains 23 codons (7.5%) which are rare in bacteria, similarly, out of a total of 1103 codons of human UPS7, 77 codons (6.9%) are rare in the *E. coli* genome. The presence of these rare codons may interrupt the protein synthesis prematurely resulting in the production of small proteins which were observed in the comassie blue stained gels and western blots of the respective fractions. To address this possibility, the Rosetta DE3 strain of *E.coli* was used to express the recombinant proteins. In both cases a minor improvement was observed in expression and procurement of the full length recombinant proteins. This observation is consistent with an earlier report, based on the assessment of 68 recombinant proteins, illustrating that recombinant proteins that express well in BL21(DE3) did not show any significant improvement in production when expressed in the Rosetta(DE3). Furthermore, in certain cases reduced expression of the recombinant protein was observed which otherwise expressed well in BL21(DE3) (Tegel *et al.*, 2010). This is possibly due to the extrametabolic pressure in the Rosetta(DE3), which harbours an extra plasmid to encode and compensate for rare codon tRNAs. However, in the same study (Tegel *et al.*, 2010), out of the 68 different recombinant proteins (encoded by genes with rare codons), some degree of improvement was observed in the procurement of the full length recombinant proteins (86% of the total tested for purity) when expressed in Rosetta(DE3) as compared to when the same proteins were expressed in the BL21(DE3) (74% of the total tested for purity). This suggests that the effect of Rosetta strains to compensate for rare codon usage for heterelogous expression of proteins, differ with the specific protein. Additionally, a more pronounced negative effect was observed in the heterelogous expression of recombinant proteins when the rare codons are present at the 5' end of the encoded gene (Kim *et al.*, 2006). It was also found that attaching a long N-terminal tag or fusion protein could improve the expression and recovery of full length recombinant protein (Tegel *et al.*, 2010). Alternatively, purification could be enhanced by using an expression vector that fuses the 6xHis tag at the C-terminus of the recombinant protein, hence theoretically, allowing only the full length recombinant protein product to bind with the Ni beads. The purification strategy could be further enhanced by employing dedicated tools for protein purification such as reverse phase or ion exchange column chromatography. However, an advantage of the array system is that complete purification of the protein is not required, indeed a complex mixture can be used, so long as the detecting antibodies are specific.

EBNA1-EBP2 interaction

Probing of the EBNA1 peptide array with recombinant GST tagged or His tagged EBP2 proteins showed nearly identical binding sites, binding to peptides incorporating GR1, GR2 and a new site represented by peptide 88 to 91 (BBR: Gly456-Lys460). Binding of EBP2 within GR1 and GR2 regions of EBNA1 has been previously demonstrated using the yeast two hybrid system and by co-immunoprecipitation studies (Shire *et al.*, 1999; Kapoor and Frappier, 2003; Nayyar *et al.*, 2009). Both GR regions are rich in glycine and arginine residues and are separated by a gly ala repeat of around 225 amino acids in the EBV EBNA1 sequence. The GR2 region of EBNA1 has been shown to have greater affinity for EBP2 interaction than GR1 and is critical (unlike GR1) for maintenance of the EBV episome (Nayyar *et al.*, 2009). The interaction of EBP2-His with peptides 88 to 91 (incorporated sequence Gln436-Leu485) has not been reported previously and in a 2 hybrid assay, deletion of this region along with C-terminal DNA binding region, did not show any noticeable effect on the EBNA1-EBP2 interaction (Shire *et al*., 1999). In contrast to GR1 and GR2, this region is not particularly rich in Gly and Arg residues, however, a stretch of three positively charged residues, preceded with two glycine residues (Gly456-Lys460) is present within the BBR of this site. It is possible in the eukaryotic cell this site may remain concealed in the tertiary conformation of EBNA1 but once exposed (as in an EBNA1 array) may form another interaction point between EBNA1 and EBP2. Moreover, in the preparative array peptide 91 did not show positive binding of EBP2-His. This could be due to the difference in shape of the peptide spot between the arrays, which is convex in the case of glass array and flat in the preparative arrays, giving the former more surface area for contact. However, in the three dimensional model of EBNA1, the BBR of peptide 88 and 91 (Arg451-Lys460) resides on a loop which is located proximal to the GR2 region (main EBP2 interaction site) (Figure 4.11). This suggests that the third site may be involved in binding with EBP2 *in vivo*. Sensitivity of the peptide array for the challenged protein relies on the purity and concentration of the spotted peptide (Volkmer *et al*., 2012). Given the purity of spotted peptide differs between analytical and preparative arrays, where the latter is less pure than the former, this may result in the observational differences in the binding partner interaction on the same sequence on those arrays. Additionally and/or alternatively, the possibility of difference in the secondary structure conformation of the peptides between the glass array and membrane array spot could also contribute in the observed differences. As recombinant protein expressed in the bacteria, which have less advance chaperone machinery to guide the protein folding, the recombinant protein molecules may adopt different structural conformation solely on the thermodynamic stability which in turn may produce observational variations within the same type or different type of arrays (analytical and preparative). However, these problems could be addressed by adopting rigorous control over the purity and concentration of the peptides and partner proteins. Secondly, expression of the recombinant proteins in eukaryotic cells especially of mammalian origin may also provide more control and may generate more consistent results on peptide arrays.

On EBNA1 preparative array, spots corresponding to peptide 10 (Gly46-Val70) and peptide 70 (Gly346-Arg370) showed positive binding with EBP2-His. Alanine substitution of peptide 70 has shown that modification of certain residues weakens/abolished the binding of EBP2-His. Previously it has been observed that small deletions in the EBP2 binding region did not affect the binding of EBP2 with EBNA1 nor the maintenance of its genome within the infected cells (Shire *et al*., 1999). Taken together, it is reasonable to propose that Glu359 to Glu368 (within peptide 70) may represent a candidate for the minimal binding region on EBNA1 for its interaction with EBP2. This possibility could be evaluated by generating truncated peptides of this region and probing them with EBP2. Furthermore, a reciprocal array (EBP2 array probed with EBNA1 protein) could further highlight the interacting region for EBNA1 on the binding partner (EBP2). Replacement of Ala with Asp (Ala56 of peptide 8) improved the binding whereas replacing Glu359 and Glu367 (of peptide 70) with alanine almost abolished the EBP2-His interaction with the corresponding peptide spot. Similarly, replacement of lysine residue (Lys379 of peptide 73) with alanine also improved the binding. Taken together these observations suggest the replacement of non polar or positively charged residues with negatively charged or hydrophobic amino acids enhances the binding of EBP2 to the peptides. Such substitutions may reduce the electrostatic and/or steric repulsion due to the abundance of positively charged residues (Arg) in the EBP2 binding regions. This is also consistent with the data which shows phosphorylation (which would introduce a negative charge) of Ser60, Ser62 (incorporated in peptide 10) and Ser 475 (within peptide 91) improved the binding of EBP2-His with EBNA1, despite the observation that alanine substitution of these residues in their native peptide did not make any difference in the binding of the EBP2-His.

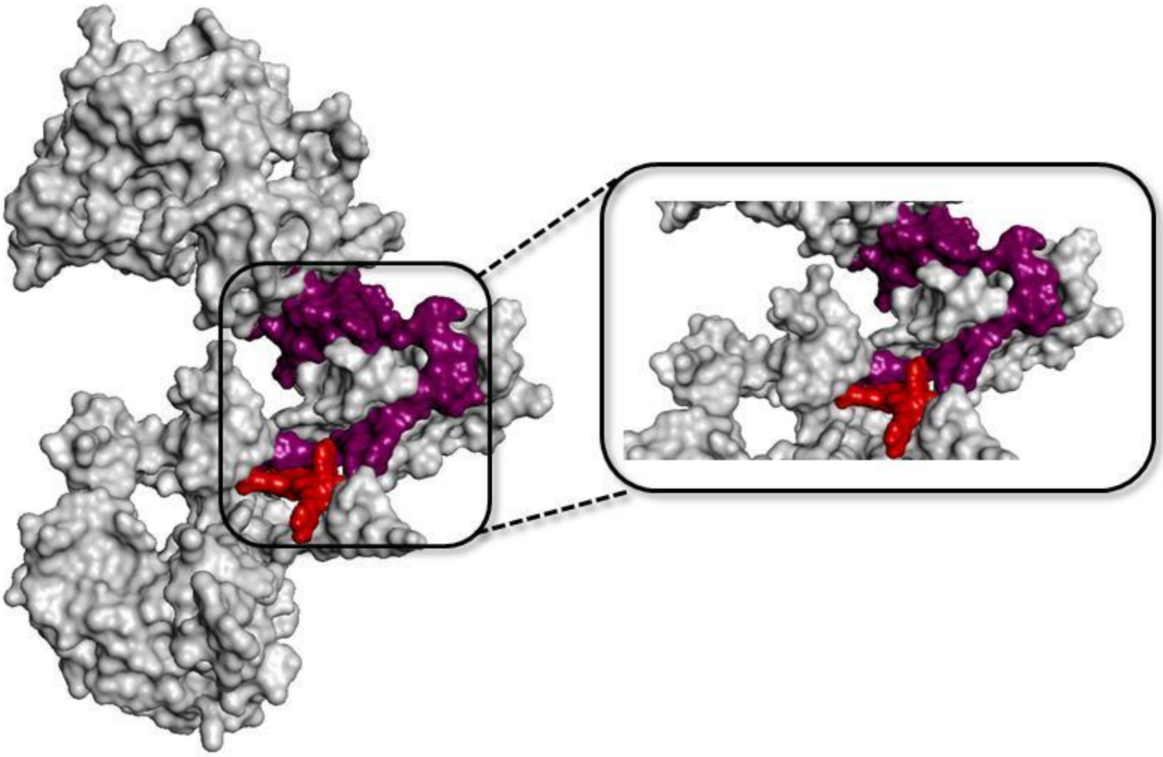


Figure 4.11. EBNA1-EBP2 interaction. Surface topology of three dimensional model of EBNA1 is shown, where the GR2 region and the BBR (Gly456-Lys460) of third site (observed in this study) are highlighted with purple and red respectively.

It is possible that these post translational modification events may not be necessary for the EBNA1-EBP2 interaction however could enhance it. However, replacement of the positively charged arginine residues in turn, with alanine did not alter binding. However, it has been shown that arginine residues, particularly those present in the GR2 region, could be methylated and this contributes to localization of EBNA1 molecules within the cells (Shire *et al*., 2006). These post translational modifications of the EBNA1 molecule are mediated by PRMT1 and PRMT5 proteins and found to alter the intracellular localization of the EBNA1 molecule (Shire *et al*., 2006). Addition of methyl groups could alter the electrostatic properties and structural features of the binding region as well as the structure and the effect of these methylations on the EBP2 interaction remains to be evaluated.

In summary, the data support previous reports in relation to the binding region of EBNA1 for EBP2 (Shire *et al*., 1999; Kapoor *et al*., 2003; Nayyar *et al*., 2009) and also highlights a few potentially critical residues involved in this interaction. The data also show the possibility that phosphorylation of serines at the N-terminus may contribute to the binding of EBP2 at the GR1 region of EBNA1. However, because of some technological limitations, validation of these observations essentially requires additional biochemical and functional studies. For example EBP2 interacts with EBNA1 in its complexed form with viral DNA, which represents the dimer state of the molecule (Shire *et al*., 1999; Kapoor *et al*., 2003; Nayyar *et al*., 2009). The peptide array of EBNA1 reflects short segments of the EBNA1 sequence. The three dimensional conformation of EBNA1 under *in vivo* conditions may make some of these sites buried and not available for the partner protein interactions. Therefore further verification, using tools for investigating the protein protein interaction, is required to substantiate or refute the present findings. Moreover, EBP2 itself can form a homodimer (Tsujii *et al*., 2000) and the recombinant EBP2 may or may not be in the dimer state. Therefore expression of EBP2 in eukaryotic cells might provide a more suitable system for producing EBP2 molecules in near native conformation.

EBNA1-USP7 interaction

Probing EBNA1 peptide arrays with USP7-His protein showed strong binding of USP7 at peptide 87, supporting previous reports and highlighting the interacting region on EBNA1 for USP7 interaction (Holowaty *et al*., 2003; Saridakis *et al*., 2005). Also consistent with the earlier studies, probing truncated peptides of EBNA1 with USP7-His demonstrated the minimal binding region spanning from Asp440 to Thr448, with Asp441, Gly 445 and Ser447 being most critical in establishing the EBNA1-USP7 interaction (Saridakis *et al*., 2005). Although in the EBNA1 octapeptide (DPGEGPST) USP7 (MATH domain) complex no hydrogen bonds were observed between Gly445 and the MATH domain of

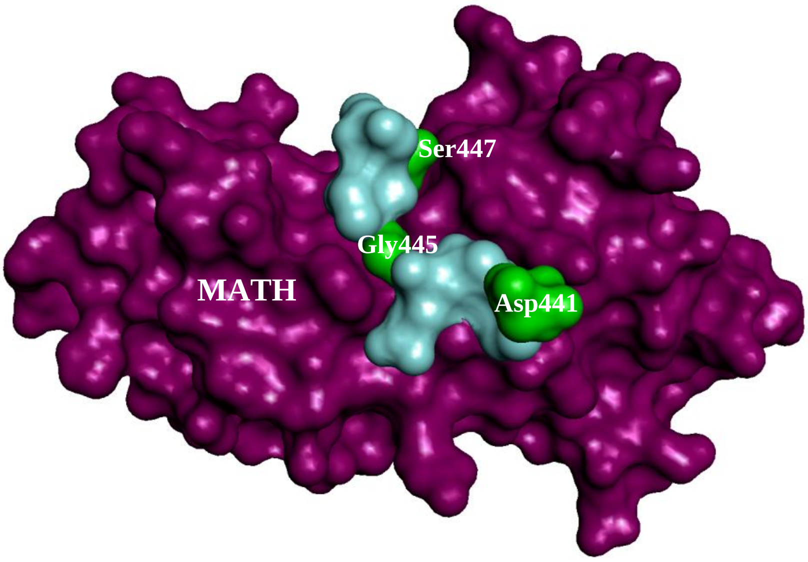
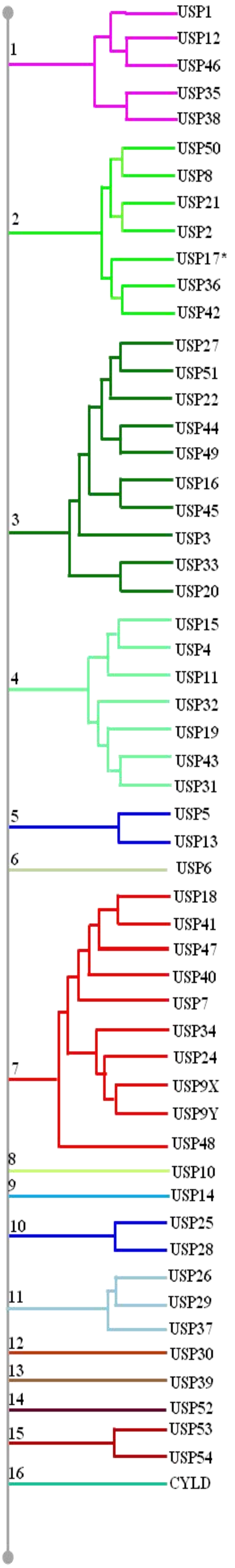
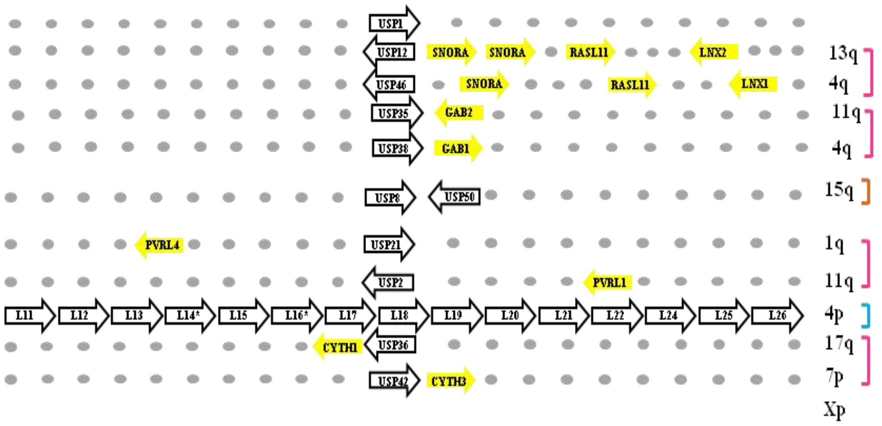


Figure 4.12. EBNA1-USP7 binding. The space filled image of EBNA1 octapeptide and USP7 MATH domain complex (PDBid; 2YY3; from Saridakis *et al*., 2005) is shown where the MATH domain is represented by purple surface while EBNA1 octapeptide is shown with cyan and green. The spatial position of the three critical EBNA1 residues (Asp441, GLy445 and Ser447) within is indicated. Note the kink in the EBNA1 peptide at the Gly445 position.

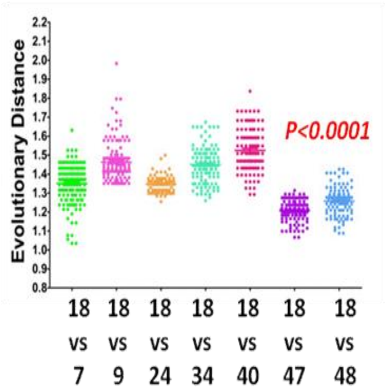
USP7, a close inspection of the atomic coordinates of this complex shows that Gly445 may provide essential flexibility for the peptide to curl and accommodate itself in the cavity of the USP7 MATH domain (Figure 4.12). This suggests that EBNA1 peptides may adopt at least some structural conformation on the array. Moreover, phosphorylation of Ser447 reduces the binding of USP7 with the EBNA1 peptide. This serine was not found to be phosphorylated in the EBNA1 extracted from the BJAB cell line (Duellman *et al*., 2009). Consistent with this, our EBNA1 structural model, Ser447 did not show a propensity to be phosphorylated.

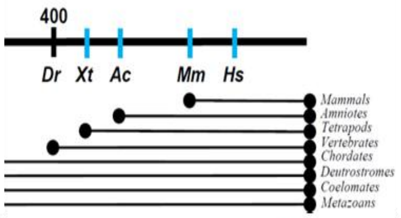
EBNA1 interacts a variety of host proteins to carry out its biological role (Frappier, 2012; Smith and Sugden, 2013). To date many of these interactions have not been mapped on the EBNA1 sequence. In addition, the binding regions of EBNA1 have also not been mapped on its respective partner proteins. Owing to the binding of anti EBNA1 antibodies and recombinant proteins to the expected protein region on EBNA1 peptide array, it is reasonable to suggest that the present approach could be reliably used to map the binding sites of other potential and known interaction partners of EBNA1. The data gathered could direct the functional assessments of those intermolecular associations and could further be exploited for designing therapeutic interventions against EBNA1.

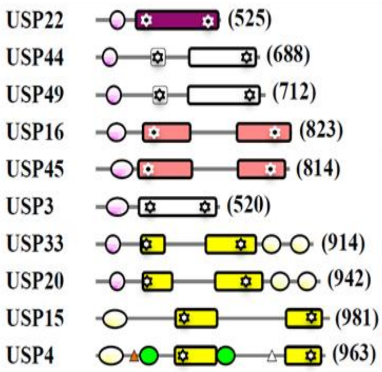


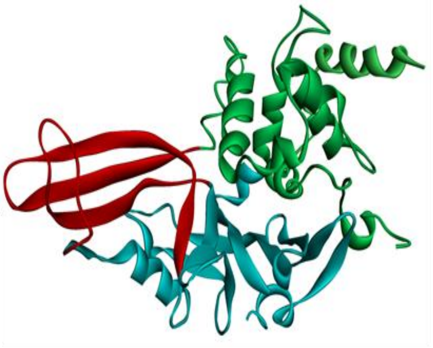


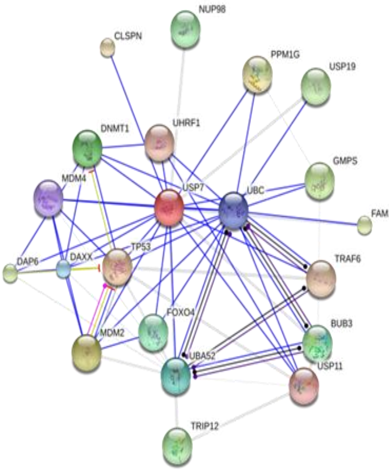
Chapter 5. Evolution of Ubiquitin Specific Protease

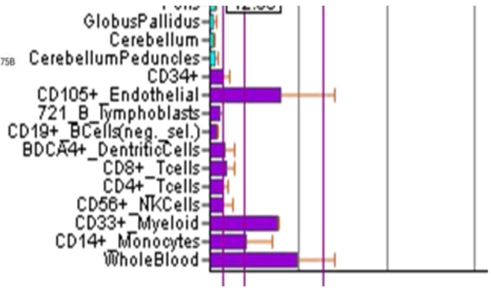












5. Results: Evolution of Ubiquitin Specific Proteases

5.1. Introduction

Ubiquitination and deubiquitination play a major role in regulating the turnover of properly and improperly folded proteins in eukaryotic cells. This chapter describes the phylogenomic analysis of a family of deubiquitinating enzymes, Ubiquitin Specific Proteases (USPs). The diversity and phylogenetic relationship of the USP homologues were explored in representative species of different taxonomic lineages of animals ranging from protozoa to primates. Species were selected on the basis of their placement in the evolutionary tree of life and availability of the full genome sequence in the public databases. In addition, to resolve any ambiguous phylogenetic relationship of USP homologues, evolutionary distances between the sequences were estimated. In order to explore the gene expansion mechanisms at work, genomic syntenies of paralogues in humans were compared. Phylogenetic analyses of USPs were also compared with their gene expression profile, protein domain composition and protein interaction network to investigate the tissue specificity, protein domain promiscuity and molecular binding partners, which in turn reflects the functional divergence of the proteins. Although some USPs of humans have been characterized extensively for their structural and functional attributes (Komander *et al.*, 2009), the evolutionary history of these genes remains largely unaddressed. To the best of my knowledge this study is the first account exploring the evolutionary history of USPs. Given the increasingly evident involvement of USPs in several human diseases including cancers, such a study holds value for both general and medical biology.

5.2. Identification of paralogous groups

The Ensembl paralogy prediction pipeline was used to collect the USP sequences encoded by the human genome. In total 86 homologues of USPs were identified in the human genome, which are divided into 16 paralogous groups (Table 5.1). Group 2 is the most densely populated, including 32 USP17 like genes (some of which are predicted as pseudogenes) in addition to USP2, USP8, USP50, USP21, USP36 and USP50. Interestingly, group 6 contains 21 paralogous genes, of which only USP6 contains a peptidase C19 domain, a characteristic domain present in all other USPs. This suggests that the paralogous relationship between USP6 and other members of group 6 (as indicated by the Ensembl) is mainly due to the presence of a TBC (Tre-2/Bub2/Cdc16) domain (found exclusively in the members of group 6).

Paralogous sets	Members
Group 1	USP1, USP12, USP35, USP38, USP46
Group 2	USP2, USP8, USP17 (32 copies including intact orf and pseudogenes), USP21, USP36, USP42, USP50
Group 3	USP3, USP16, USP20, USP22, USP33, USP44, USP45, USP49, USP51, USP27
Group 4	USP4, USP11, USP15, USP19, USP31, USP32, USP43
Group 5	USP5, USP13
Group 6*	USP6, USP6NL, TBC1D3(19X)
Group 7	USP7, USP9X, USP9Y, USP18, USP24, USP34, USP40, USP41, USP47, USP48
Group 8	USP10
Group 9	USP14
Group 10	USP25, USP28
Group 11	USP26, USP29, USP37
Group 12	USP30
Group 13	USP39
Group 14	USP52
Group 15	USP53, USP54
Group 16	CYLD

Table 5.1. Paralogous groups of Ubiquitin Specific Peptidase. Shown here is the distribution of human USPs in 16 paralogous groups, as predicted by the Ensembl genome browser paralogy pipeline. Asterisk (*) denotes that except USP6 other homologues present in this group do not contain the characteristic peptidase C19 domain, which is present in all other USPs. The homologous/paralogous relationship between USP6 and other members of this group is thus based on the presence of a TBC domain.

5.3. Distribution of USPs across the animal kingdom

To investigate the distribution of USPs across the animal kingdom, protein sequences of full length human USP homologues and their corresponding C19 domain were BLASTed against the genomes of selected organisms (Table 5.2). Organisms were selected that are located at important transitional and/or speciation events in the evolutionary tree. *Homo sapiens* (Hs, human), *Mus musculus* (Mm, mouse), *Bos taurus* (Bt, cow) and *Canis lupus familaris* (Cf, dog) genomes were screened to represent the major lineages of eutherian mammals; representing primates, rodentia, cetartiodactyla and carnivora, respectively. Genomes of *Monodelphis domestica* (Md, opossum) and *Ornithorhynchus ananitus* (Oa, platypus) were explored as representatives of early divergent mammalian lineages, those of marsupial mammals and monotremata respectively. To represent the non mammalian vertebrates: reptiles, amphibians and fishes, genomes of *Anolis carolinesis* (Ac, lizard), *Xenopus tropicalis* (Xt, clawed toad) and *Danio rerio* (Dr, zebra fish), respectively, were screened. *Ciona intestinalis* (Ci, tunicate) was taken as a model for non vertebrate chordates. *Strongylocentrotus purpuratus* (Sp, sea urchin) reflects a non chordate deuterostome lineage which separated from the common ancestor with all other chordates

Grps	USP	Hs	Mm	Bt	Cf	Md	Oa	Ac	Xt	Dr	Ci	Sp	Dm	Ce	Hm	Dd	Cr
1	1	🟫	🟫	🟫	🟫	🟫		🟫	🟫	🟫		🟫	🟫				
	12	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫			🟫	🟫
	46	🟫	🟫	🟫	🟫	🟫	🟫	🟫		🟫							
	35	🟫	🟫	🟫	🟫	🟫		🟫	🟫	Lc	🟫	🟫	🟫		🟫	🟩	
	38	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫							
2	2	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟩	🟫		🟩		
	21	🟫	🟫	🟫	🟫	🟫		🟫	🟫	🟫							
	17	🟫	🟫	🟫	🟫												
	36	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫		🟫	🟫	🟫
	42	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫							
	8	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟩	🟫		🟩
	50	🟫	🟫	🟫	🟫	🟫	🟫	🟫		Ga							
3	3	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫		🟫		🟫	🟫		
	16	🟫	🟫	🟫	🟫	🟫		🟫	🟫	🟫	🟫	🟫	🟫		🟫	🟫	
	45	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫							
	20	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟩	🟫	🟫	🟩	🟩		
	33	🟫	🟫	🟫	🟫	Sh	🟩	🟫	🟫	🟫							
	22	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟩		
	27	🟫	🟫	🟫	🟫												
	51	🟫	🟫	Ss													
	44	🟫	🟫	🟫	🟫	🟫	🟩	🟫	🟫	🟫	🟫	🟩					
	49	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫							
4	4	🟫	🟫	🟫	🟫	🟫	🟩	🟫	🟫	🟫	🟫	🟫		🟫	🟫	2	🟫
	15	🟫	🟫	🟫	🟫	🟫	🟩	🟫	🟫	Ga							
	11	🟫	🟫	🟫	🟫	🟫	🟫	🟫		🟫							
	31	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫				
	43	🟫	🟫	🟫	🟫	🟫		🟫	🟫	🟫							
	19	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫		🟩	🟫		
	32	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫		🟩	🟫	
5	5	🟫	🟫	🟫	🟫	🟫		🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫
	13	🟫	🟫	🟫	🟫	🟫		🟫	🟫	🟫							
6	6	🟫															
7	7	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫
	9X	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	Ga	🟫	🟫	🟫		🟫		
	9Y	🟫	🟫	🟫	🟩												
	24	🟫	🟫	🟫	🟫	🟫		🟫	🟫	🟫	🟫			🟩	🟫		
	34	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫		🟫	🟫		🟫	2	
	41	🟫															
	18	🟫	🟫	🟫	🟫	🟫		🟫	🟫	🟫							
	48	🟫	🟫	🟫	🟫	🟫	🟩	🟫	🟫	🟫	🟫	🟫		🟩	🟫	🟫	🟫
	40	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫				🟫	
	47	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	Ga	🟫	🟫	🟫	🟫	🟫		
8	10	🟫	🟫	🟫	🟫	Sh	🟫	🟩	🟫	🟫	🟫	🟫	🟫		🟫	🟫	🟫
9	14	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫
10	25	🟫	🟫	🟫	🟫	Sh	🟫	🟫	🟫	🟫	🟫	🟫		🟫			
	28	🟫	🟫	🟫	🟫	🟫	🟩	🟫	🟫	🟫							
11	26	🟫	🟫	🟫	🟫												
	29	🟫	🟫	Bm													
	37	🟫	🟫	🟫	🟫	🟫	🟩	🟫	🟫	🟫	🟫	🟫			🟫		
12	30	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫			
13	39	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫
14	52	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫
15	53	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫				
	54	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫	🟫							
16	CYLD	🟫	🟫	🟫	🟫	🟫	🟫		🟫	🟫	🟫	🟫	🟫	🟫			
Total		55	53	53	51	47	38	46	44	47	29	31	23	17	23	17	11

Table 5.2. Distribution of USPs across the animal kingdom. The presence in the given genome of USP homologues, as identified by genomic BLAST of selected animals is indicated (shaded brown). The incomplete sequences with unclear orthologous relationship (hence not included in the analysis) are shaded green. The total number of genes (including USP homologues with C19 domains and excluding the multiple homologues of USP17) found in particular species are indicated in the last row. Also excluded are the plant USP homologues found in Dd and Cr. Genes detected in species (other than the species abbreviated in the row header, as indicated in the section 1.3) of the same taxonomic/evolutionary group are shown: *B. mutus* (Bm), *L. chalumenae* (Lc), *S. harissi* (Sh) and *G. aculateus* (Ga). The lack of clear orthology with human/vertebrate USP genes (probably reflecting the orthologues prior to duplication in vertebrate species) are represented by fused boxes.

around 740 million years ago. *Drosophila melanogaster* (Dm, fruit fly) and *Caenorhabditis elegans* (Ce, nematode) were used as representatives of protostomes. *Hydra manipulata* (Hm, hydra) and *Dictyostelium discodium* (Dd, slime mould) genomes provide information about the metazoa-protozoan split, which occurred more than 1100 million years ago (MYA). While the genome of green algae, *Chlamydomonas reinhardtii* (Cr, chlamydomonas) was screened for USP homologues. Collectively with slime mould, it provides the information for the presence of USP homologues before plant-animal split (>1200 MYA). Where the orthologues could not be found in these species, BLAST search was extended to another species of the corresponding taxonomic lineage in order to generalize the observations. For example, the mouse, opossum, cow and dog genome BLAST search was expanded to *Rattus rattus* (Rr, rat), *Sarcophilus harissi* (Sh, Tasmanian devil) *Sus scrofa* (Ss, pig) and *Ailuropoda melanoleuca* (Am, panda) respectively. Similarly, *Gasterosteus aculeatus* (Ga, stickleback) and *Latimera chalumenae* (Lc, Coelacanth) were generally used for the zebra fish and *Brachiostoma floridae* (Bf, lancelet) genome was screened in place of tunicate genomes respectively. Only genes with the characteristic peptidase C19 domain were included in the analysis. Only one of the multiple homologues of USP17 (only found in eutherian mammals) was included to better reflect the long term evolutionary changes in the number of USP homologues across the animal kingdom (Table 5.2).

In total, 55 paralogues of USPs (only including genes with the peptidase C19 domain), encoded by the human genome, were identified. With the exception of platypus, most of these homologues were traced back to zebra fish (Table 5.2). Two homologues (USP6 and USP41) were found only in humans and great apes (primate lineage). Several genes in vertebrates are annotated as USP6-N-terminal like (USP6NL), primarily due to the presence of a TBC domain at the N-terminus, but do not have a peptidase C19 domain. USP29 and USP51 orthologues were only found in humans, mice and cow/pig (of the species examined). BLAST results show that USP17, USP26 and USP27X homologues are restricted to eutherian mammals. The platypus genome BLAST shows the existence few USP homologues compared to human and zebra fish, indicative of extensive species/lineage specific gene loss in monotremes or incomplete genome assembly and/or annotation. Similarly, certain USP homologues (including USP46 and USP50, present in bony fishes and mammals) were not identified in the genome of clawed toad suggesting species/lineage specific gene loss.

Taken together, the genome BLAST results demonstrate that the existing array of vertebrate USPs share deep evolutionary history, as orthologues of almost all of the human USPs exist in non mammalian vertebrates, including fish. Additionally, the total number of

USP homologues points to an extensive expansion of USPs in early vertebrates (Table 5.2). In most cases, at least one paralogue of each group is present in protozoa. This suggests that the USP paralogous groups share a deep ancestral history which predates the protozoa-metazoan split. In order to explore the USP homologues in more distantly related lineages, the BLAST search was extended to chlamydomonas and microbial genomes. The algae chlamydomonas is a modern day representative of the ancestral root of plants and BLAST search revealed the presence of 13 USP homologues, of which 11 showed homology with the USPs found in different animal species. This suggests the presence of USP proteins in the common ancestor of plants and animals. A BLAST search using the C19 protein domain sequence of all available microbial genomes in the databases provided only two positive hits (nearly identical homologues) in the bacterial species *Candidatus amoebophilus asiaticus* of phylum bacteriodes, indicative of a prokaryotic origin of USPs. However, given the breadth of microbial genomes examined, horizontal gene transfer offers a likely explanation for the presence of the peptidase C19 domain in microbes.

5.4. Phylogenetic analyses of USPs

The robustness of phylogenetic reconstruction of a gene family is heavily dependent on the accuracy of sequence alignment. As observed in the datamining, USP homologues in humans and other animals vary considerably in protein length and protein domain architecture. Thus in order to reconstruct the phylogenetic history of USPs, only the common peptidase C19 domain protein sequences were aligned. Similarly, to optimize the reliability of the tree, phylogenetic trees of each paralogous group were reconstructed separately from homologues of both vertebrates and non vertebrates (NV, including invertebrates and non vertebrate chordates) using the maximum likelihood method. NV homologues were included/excluded in the tree construction on the basis of BLAST and reciprocal BLAST score to the query sequence (human USP homologues). For reconstruction of evolutionary trees for each paralogous group, only those NV homologues were included which showed first ranked similarity (highest blast score and lowest e value) with any gene included in the respective paralogous group, in both BLAST and reciprocal BLAST. Trees were reconstructed with and without rooting with the C19 domain sequence of a protein from bacteria *Candidatus amoebophilus asiaticus* (YP003573189), one of the two (nearly identical) C19 domain bearing prokaryotic proteins. For clarity, the evolutionary relationship is described separately for each paralogous group.

Group 1. Group 1 comprises five USP homologues: USP1, USP12, USP35, USP38 and USP46. In phylogenetic tree reconstruction, these paralogues form distinct clades of USP12/46, USP35/38 and USP1 (Figure 5.1).

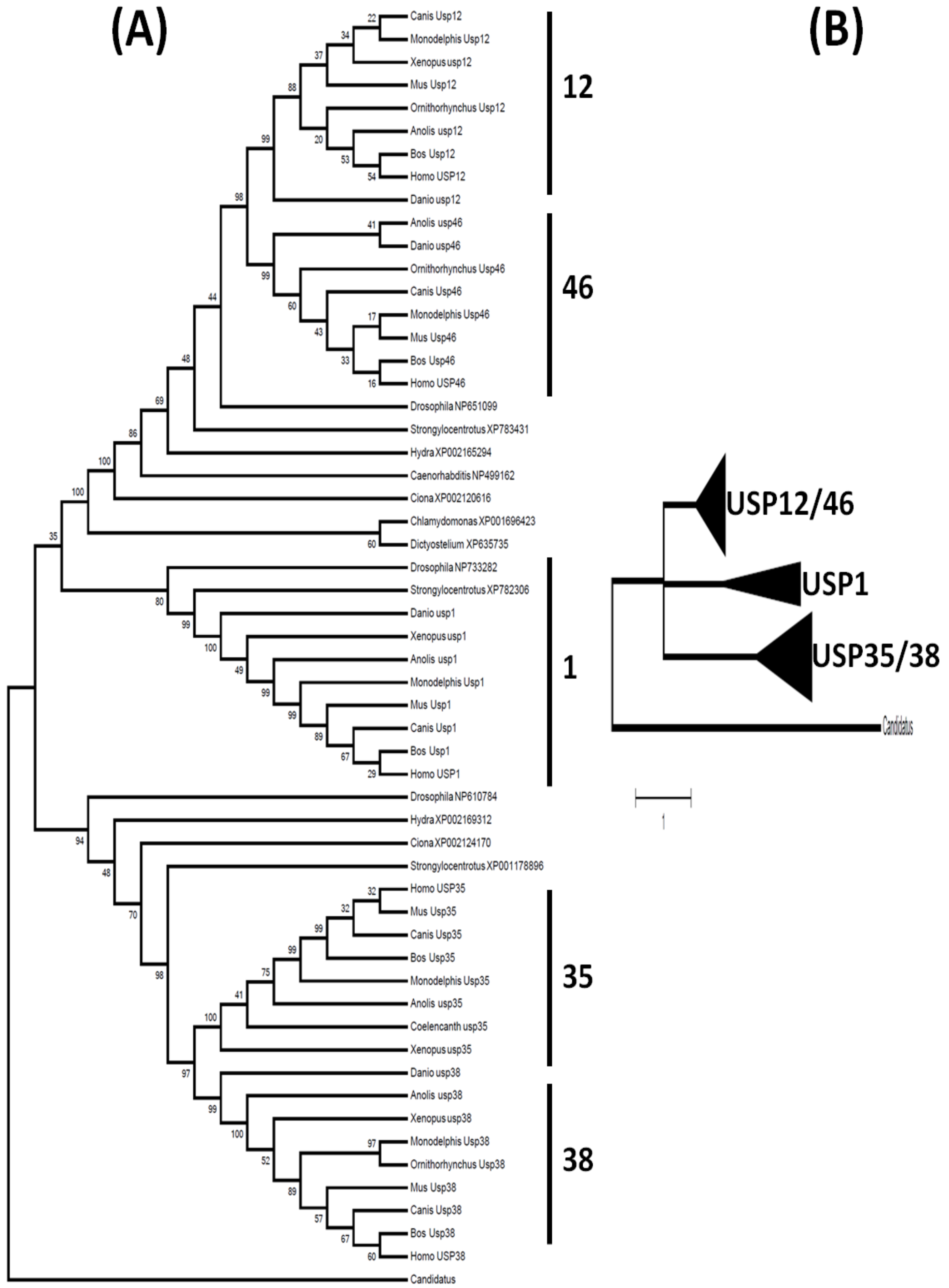


Figure 5.1. Phylogenetic history of group 1 USP homologues. (A) A phylogenetic tree was reconstructed by the maximum likelihood method using C19 domain protein sequences, adopting the Jones Taylor and Thornton model with gamma distribution and with some invariable sites (JTT+G+I) (Jones *et al.*, 1992). All vertebrate homologues are annotated according to their orthologous relationship, whereas NV homologues are represented by the accession numbers. **(B)** Branch length format of the same tree is shown where all monophyletic clades are collapsed as indicated. Note the difference in the evolutionary rate between different branches.

Distinct homologue of this group was identified each in slime mould and chlamydomonas which shows orthology with the USP12/46 clade (100% bootstrap support) while other partial sequence of slime mould (XP645628) (found by the genome BLAST) shows similarity with the USP35/38 clade. This suggests a premetazoan origin of USP12/46 and USP35/38 ancestral genes. Single NV homologues (of several species examined) of USP12/46 and USP35/38 outgroup the two vertebrate specific subclades USP12 and USP46, USP35 and USP38 respectively, reflecting the expansion of these USPs prior to vertebrate divergence (parallel with timing of whole genome duplication events (1R and/or 2R)) (Figure 5.1A). USP12/46 clade share ancestry with the USP1 clade which origin could be reliably placed at least with the emergence of coelomates, owing to the presence of fruit fly and sea urchin homologues in the respective clade. However, as it outgroup USP12/46 clade, it is possible that USP1 also has a premetazoan origin, but had lost in slime mould and hydra. Alternatively, as apparent from the branch lengths of USP1 orthologues, the higher substitution rate may result in the outgrouping of this clade (Figure 5.1B).

Group 2. Group 2 includes seven USP paralogues (discounting the multiple homologues of USP17): USP2, USP8, USP17, USP21, USP36, USP42 and USP50. Phylogenetic reconstruction of the homologues reveals distinct clades of USP8/USP50, USP2/21 and USP17/36/42 (Figure 5.2).

In genome BLAST, two homologues of this group were identified in chlamydomonas while one homologue was found in the slime mould, all groups with USP17/36/42 clade, reflecting the origin of this group at or before plant-animal divergence. Two homologues of hydra show orthologous relationships with USP8 (sharing ancestry with USP50) and USP17/36/42 clades and one partial sequence (XP004208370) of the organism shows similarity with USP2/21 clade. This suggests the divergence of USP8, USP17/36/42 and USP2/21 ancestral gene occurred in or at the common ancestor of metazoans. Interestingly, a fruit fly homologue joins the vertebrate USP2 subclade with relatively weak bootstrap support (44%) while a tunicate homologue forms an outgroup branch to both USP2 and USP21 subclades forming a single monophyletic clade (97% bootstrap support) (Figure 5.2). The slightly off position of the fruit fly or tunicate sequence could be explained in terms of difference in substitution rate in either gene. Single NV homologues (except fruit fly NP608462) out group vertebrate specific USP2 and USP21 subclades, USP36 and USP42 sub clades, similarly the USP50 clade is constituted by only vertebrate homologues. This suggests expansion of these USPs with the origin of vertebrates. USP17 homologues were only identified in eutherian mammals (of those examined) suggesting a recent origin of the genes despite outgrouping all the chordates USP36/42 homologues

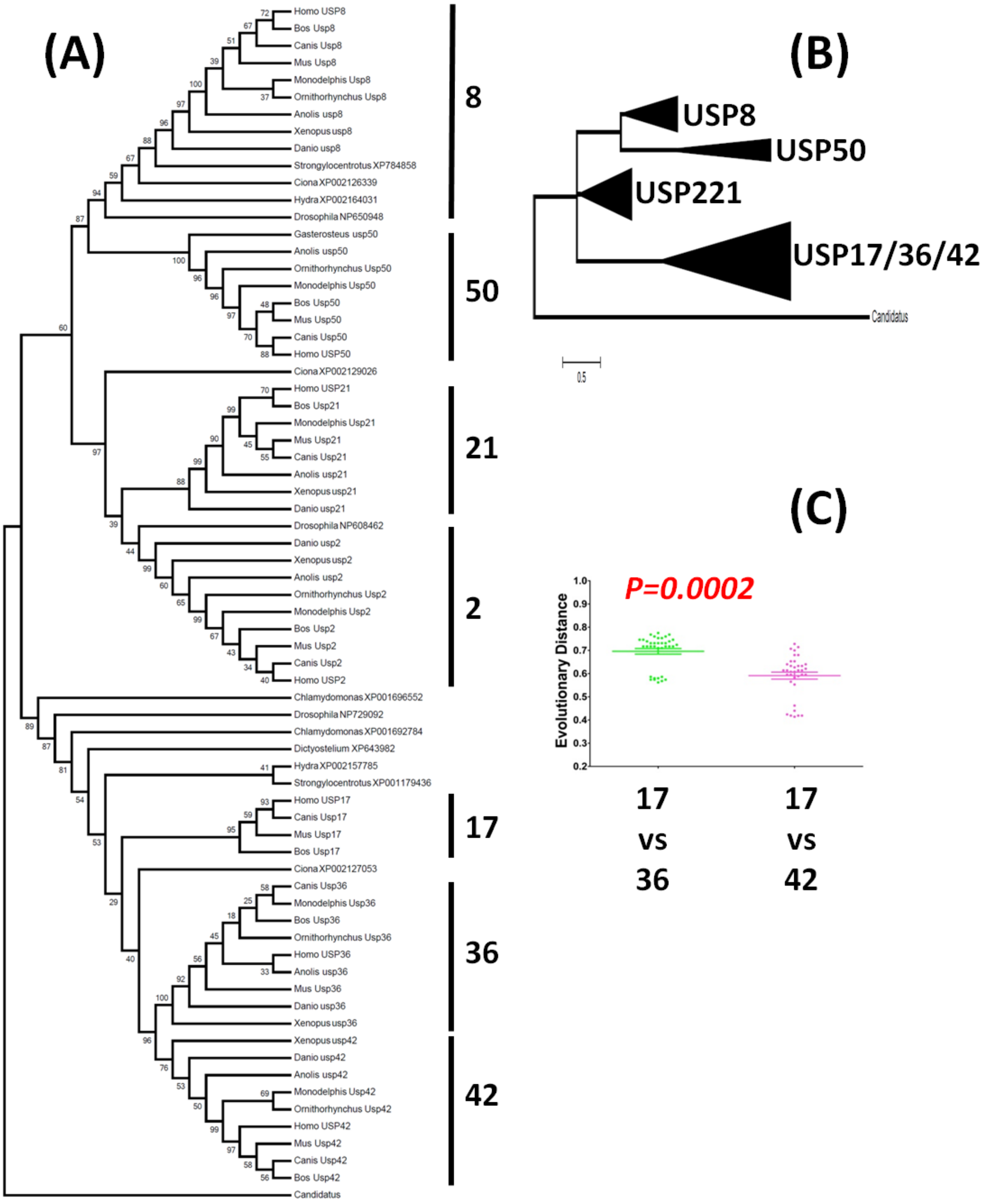


Figure 5.2. Phylogenetic history of group 2 USP homologues. (A) A phylogenetic tree was reconstructed by the maximum likelihood method using C19 domain protein sequences, adopting the Jones Taylor and Thornton model with Gamma distribution and with some invariable sites (JTT+G+I) (Jones *et al.*, 1992). All vertebrate homologues are annotated according to their orthologous relationship, whereas NV homologues are represented by the accession numbers. **(B)** Branch length format of the same tree is shown where all monophyletic clades are collapsed as indicated. Differences in the branch length represent noticeable change in the substitution rate. **(C)** Graph shows the pair wise evolutionary distance between USP17 and USP36 (green) or USP17 and USP42 (pink). Large and small horizontal bars represent mean and standard error of mean respectively. The statistical significance of the difference in evolutionary distance was estimated by Wilxon signed rank test and the estimate of significance is indicated suggesting that USP17 is similar to USP42.

(Figure 5.2A), therefore contradicts the species evolutionary tree. The branch length of USP36/42/17 clade indicate faster evolution compared to other monophyletic clades of group 2 (Figure 5.2B), nevertheless the pair wise evolutionary distance comparison suggest the similarity between USP17 and USP42 (Figure 5.2C). With regard to USP17, there are 32 homologues (nearly identical) encoded by the human genome and in a previous phylogenetic analysis of the USP17 subfamily, paralogues genes from each species form distinct clades from orthologues, which reflects intraspecies expansion of USP17 in different eutherian mammals (Burrows *et al.*, 2010). For convenience, the C19 domain sequence of only the largest protein encoding homologue of USP17 from each species (where found) is considered in this analysis.

Group 3. Group 3 comprises ten USP paralogues: USP3, USP16, USP20, USP22, USP27, USP33, USP44, USP45, USP49 and USP51. These paralogues form distinct clades in the tree with USP27 and USP51 grouping with USP22, a USP44/49 clade, a USP16/45 clade and a USP33/20 clade sharing ancestry with USP3 clade (Figure 5.3).

One distinct homologue of slime mould was identified in this group which outgroups the USP16/45 clade, placing the origin of this group in the common ancestor of animalia. Four USP homologues of group 3 were identified in hydra, of these one aligns with USP22/51/27 clade which shares ancestry with USP44/49 clade, two others with USP3 and USP16/45, whereas one partial sequence (XP004206986) shows similarity with USP20/33 clade (Figure 5.3). This suggests that the ancestral gene of group 3 had expanded to four genes at least by the emergence of the common ancestor of metazoans. These genes may have been present in the protozoan ancestor but since may have been lost from slime mould. The next evidence for expansion is the appearance of tunicate homologue of USP44/49 clade. Separation of two subclades is evident at the root of vertebrates for USP16 and USP45; USP44 and USP49 and USP33 and USP20. USP27 and USP51 homologues were only found in some eutherian mammals suggesting a recent separation from USP22 (Figure 5.3).

Group 4. Group 4 includes seven USP paralogues: USP4, USP11, USP15, USP19, USP31, USP32 and USP43. In the evolutionary tree USP4/11/15 group together to form a single monophyletic clade, USP31/43 forms a clade, whereas USP19 and USP32 individually forms two distinct clades (Figure 5.4).

In chlamydomonas genome, only one homologue of group 4 USP was found which groups with USP4/5/11 clade suggesting the origin of this group before or at the common ancestor of animal and plant. Whereas two homologues of slime mould align with USP4/11/15 clade, and one groups with USP32 clade, reflecting the premetazoan divergence within group 4. One hydra homologue shares shows lineal ancestry with USP4/11/15 clade

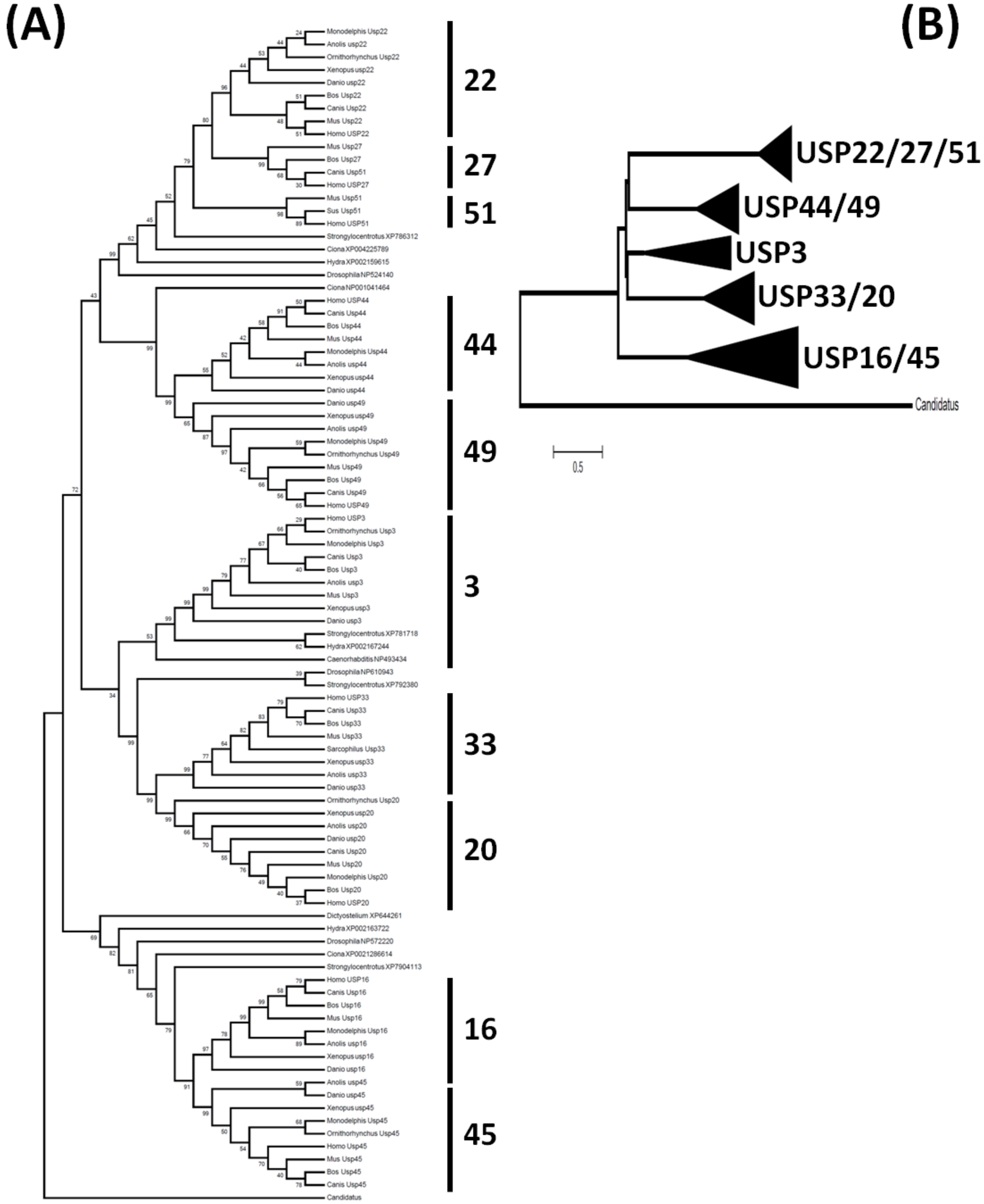


Figure 5.3. Phylogenetic history of group 3 USP homologues. (A) A phylogenetic tree was reconstructed by the maximum likelihood method using C19 domain protein sequences, adopting the Jones Taylor and Thornton model with Gamma distribution and with some invariable sites (JTT+G+I) (Jones *et al.*, 1992). All vertebrate homologues are annotated according to their orthologous relationship as found in studies, whereas NV homologues are represented by the accession numbers. **(B)** Branch length format of the same tree is shown; note the difference in the evolutionary rate (substitution rate) in different monophyletic clades.

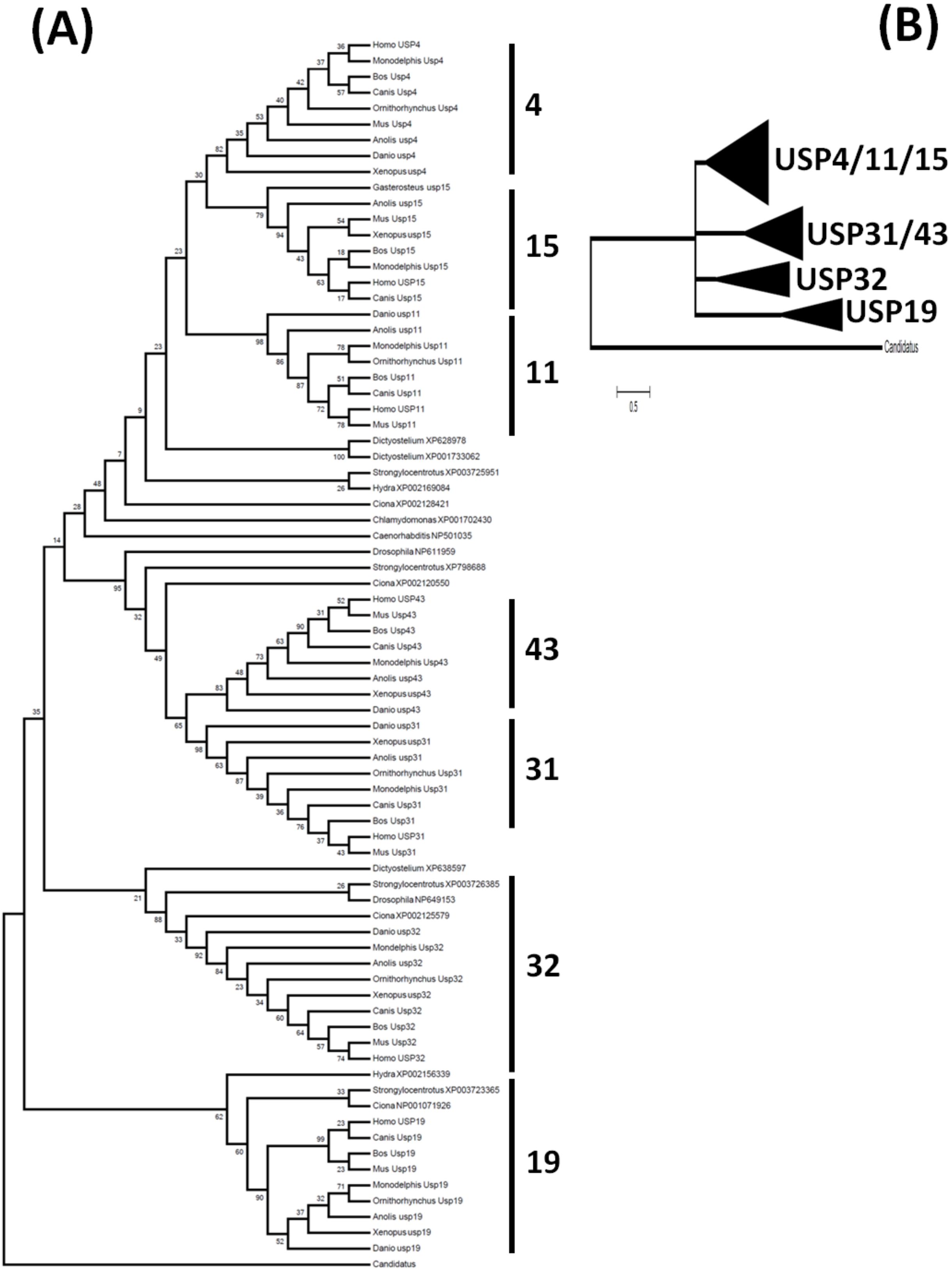


Figure 5.4. Phylogenetic history of group 4 USP homologues. (A) A phylogenetic tree was reconstructed by the maximum likelihood method using C19 domain protein sequences, adopting the Jones Taylor and Thornton model with Gamma distribution (Jones *et al*., 1992). All vertebrate homologues are annotated according to their orthologous relationship, whereas NV homologues are represented by the accession numbers. **(B)** Branch length format of the same tree where all monophyletic clades are collapsed is shown.

whereas one groups with USP19 clade. A single partial sequence of hydra (XP00216354564) shows similarity with USP32 homologues in the BLAST search. This suggests that early expansion of group 4 USPs occurred prior to the divergence of metazoan resulting in the formation of USP4/11/15, USP32 and USP19 ancestral genes. Subsequent expansion in this group was observed in USP31/43 with orthologues present in vertebrates, tunicate, sea urchin and fruit fly. Single NV homologues (except slime mould) outgroup the vertebrate specific USP4, USP11 and USP15 subclades; USP31 and USP43 subclades, indicative of vertebrate specific expansion of USPs (Figure 5.4).

Group 5. Group 5 comprises two USP paralogues: USP5 and USP13. Vertebrates USP5 and USP13 homologues form separate subclades in the phylogenetic tree and share lineal ancestry with one homologue of tunicate, sea urchin, fruit fly, nematode, hydra, slime mould and chlamydomonas (Figure 5.5A). This suggests the existence of a single ancestral gene in the common ancestor of plant and animals, which expanded into separate USP5 and USP13 genes at the base of vertebrates. The branch length of both vertebrate specific USP5 and USP13 shows that since divergence both genes have evolved at different rate (Figure 5.5B).

Group 7. Group 7 comprises nine USP paralogues: USP7, USP9 (X and Y), USP18, USP24, USP34, USP40, USP41, USP47 and USP41. All of these paralogues (except USP41) form distinct clades in the phylogenetic tree reconstruction (Figure 5.6A).

BLAST results revealed existence of two USP homologues in the chlamydomonas showing homology with group 7 members. In phylogenetic reconstruction these homologues group with USP7 and USP48, suggesting pre animal-plant split origin and divergence of the group. In slime moulds five homologues of this group were identified which aligned with the USP7, USP34 (two homologues) USP40 and USP48, reflecting the expansion of USPs of group 7 prior to the origin of metazoans. Six USP homologues of hydra show similarity with group 7 paralogues and in the phylogenetic analysis, of which two align with USP7 and USP48. The remaining four show lineal ancestry with USP47 (showing ancestry with USP40), USP34 (sharing ancestry with USP9 and USP24), USP24 and USP9. A distinct clade of USP18/41 is populated with only vertebrates homologues representing the vertebrate specific expansion of the gene (Figure 5.6), however it out groups all other monophyletic clades of group 7, potentially due to the difference in the substitution rate (Figure 5.6B). The pairwise evolutionary distance between USP18/41 clade and other paralogues of the group 7 was compared (Figure5.6C). In this comparison USP18 shows least distance to USP47 compared to any other USP of group7 and the difference between the least two (USP47 and USP48) is statistically significant ($p<0.0001$). This suggests that USP18 may have originated from USP47 but then diverged

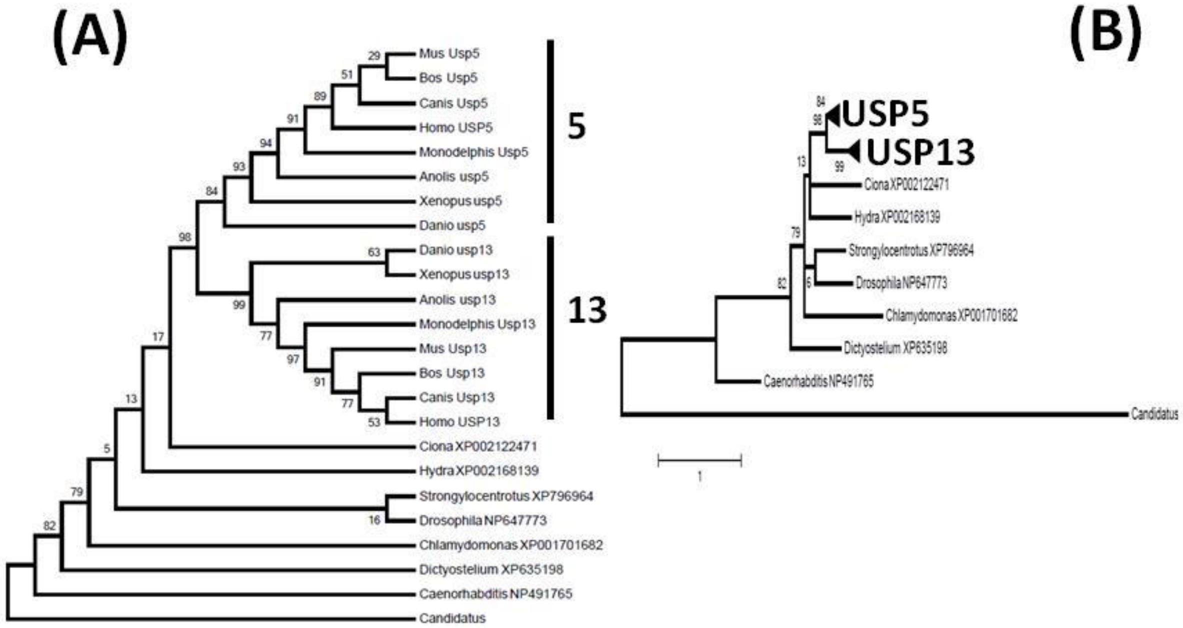


Figure 5.5. Phylogenetic history of group 5 USP homologues. (A) A phylogenetic tree was reconstructed by the maximum likelihood method using C19 domain protein sequences, adopting the Jones Taylor and Thornton model with Gamma distribution and with some invariable sites (JTT+G+I) (Jones *et al*., 1992). All vertebrate homologues are annotated according to their orthologous relationship, whereas NV homologues are represented by the accession numbers. **(B)** Branch length format of the same tree where all monophyletic clades are collapsed as indicated.

at different rate resulting in its off positioning in the phylogenetic reconstruction. USP9Y homologues share close ancestry with USP9X and only found in the eutherian mammals, indicative of recent expansion in the group 7 of USP. A USP homologue, USP41 found only in the humans and chimpanzee and clade with USP18 providing another evidence of recent expansion in the group 7 USP.

Group 10. Group 10 includes two USP paralogues: USP25 and USP28. The evolutionary tree separates the vertebrates orthologues of USP25 and USP28 into separate subclades which share lineal ancestry with one homologue of tunicate, sea urchin and nematode worm pointing to the origin of the subclades before the divergence of bony fishes (Figure 5.7).

Group 11. Group 11 is composed of three USP paralogues: USP26, USP29 and USP37. The USP37 subclade includes all the compared vertebrate homologues and shares common ancestry with the subclades of USP26 and USP29, which include of eutherian mammals and of rodentia and primates respectively. This places the origin of USP26 and USP29 from USP37 with the emergence of eutherian mammals and their separation at the common ancestor of rodentia and primates. All three subclades share a lineal ancestry with three NVs homologues of tunicate, sea urchin and hydra (Figure 5.12).

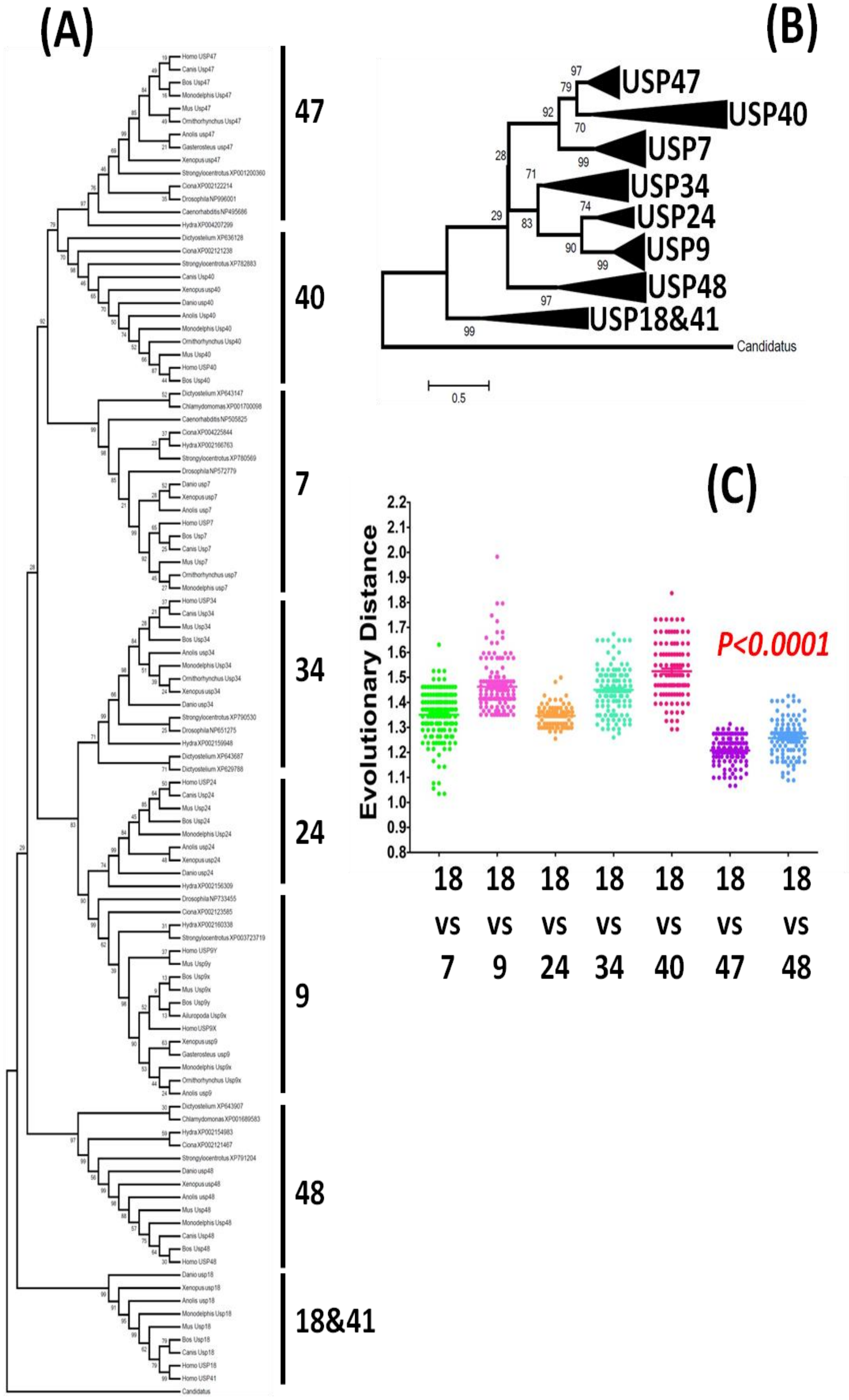

(A)
Homo USP47
Canis Usp47
Bos Usp47
Monodelphis Usp47
Mus Usp47
Ornithorhynchus Usp47
Anolis usp47
Gasterosteus usp47
Xenopus usp47
Strongylocentrotus XP001200360
Ciona XP002122214
Drosophila NP996001
Caenorhabditis NP495686
Hydra XP004207299
Dictyostelium XP636128
Ciona XP002121238
Strongylocentrotus XP782883
Canis Usp40
Xenopus usp40
Danio usp40
Anolis Usp40
Monodelphis Usp40
Ornithorhynchus Usp40
Mus Usp40
Homo USP40
Bos Usp40
Dictyostelium XP643147
Chlamydomonas XP001700098
Caenorhabditis NP505825
Ciona XP004225844
Hydra XP002166763
Strongylocentrotus XP780569
Drosophila NP572779
Danio usp7
Xenopus usp7
Anolis usp7
Homo USP7
Bos Usp7
Canis Usp7
Mus Usp7
Ornithorhynchus usp7
Monodelphis usp7
Homo USP34
Canis Usp34
Mus Usp34
Bos Usp34
Anolis usp34
Monodelphis Usp34
Ornithorhynchus Usp34
Xenopus usp34
Danio usp34
Strongylocentrotus XP790530
Drosophila NP651275
Hydra XP002159948
Dictyostelium XP643687
Dictyostelium XP629788
Homo USP24
Canis Usp24
Mus Usp24
Bos Usp24
Monodelphis Usp24
Anolis usp24
Xenopus usp24
Danio usp24
Hydra XP002156309
Drosophila NP733455
Ciona XP002123585
Hydra XP002160338
Strongylocentrotus XP003723719
Homo USP9Y
Mus Usp9y
Bos Usp9x
Mus Usp9x
Bos Usp9y
Ailuropoda Usp9x
Homo USP9X
Xenopus usp9
Gasterosteus usp9
Monodelphis Usp9x
Ornithorhynchus Usp9x
Anolis usp9
Dictyostelium XP643907
Chlamydomonas XP001689583
Hydra XP002154983
Ciona XP002121467
Strongylocentrotus XP791204
Danio usp48
Xenopus usp48
Anolis usp48
Mus Usp48
Monodelphis Usp48
Canis Usp48
Bos Usp48
Homo USP48
Danio usp18
Xenopus usp18
Anolis usp18
Monodelphis Usp18
Mus Usp18
Bos Usp18
Canis Usp18
Homo USP18
Homo USP41
Candidatus
47
40
7
34
24
9
48
18&41
(B)
USP47
USP40
USP7
USP34
USP24
USP9
USP48
USP18&41
Candidatus
97
79
92
70
99
28
71
74
29
83
90
99
97
99
0.5
(C)
Evolutionary Distance
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
P<0.0001
18 vs 7
18 vs 9
18 vs 24
18 vs 34
18 vs 40
18 vs 47
18 vs 48

Figure 5.6. Phylogenetic history of group 7 USP homologues. (A) A phylogenetic tree was reconstructed by the maximum likelihood method using C19 domain protein sequences, adopting the Jones Taylor and Thornton model with gamma distribution and with some invariable sites (JTT+G+I) (Jones *et al.*, 1992). All vertebrate homologues are annotated according to their orthologous relationship as found in studies, whereas NV homologues are represented by the accession numbers. **(B)** Branch length format of the same tree is shown where all monophyletic clades are collapsed as indicated. **(C)** Graph shows the pair wise evolutionary distance between USP18 and other members of the group 7. The difference between the least two evolutionary distances ((USP18 and USP47) and (USP18 and USP48)) suggests that USP18 is similar to USP47.

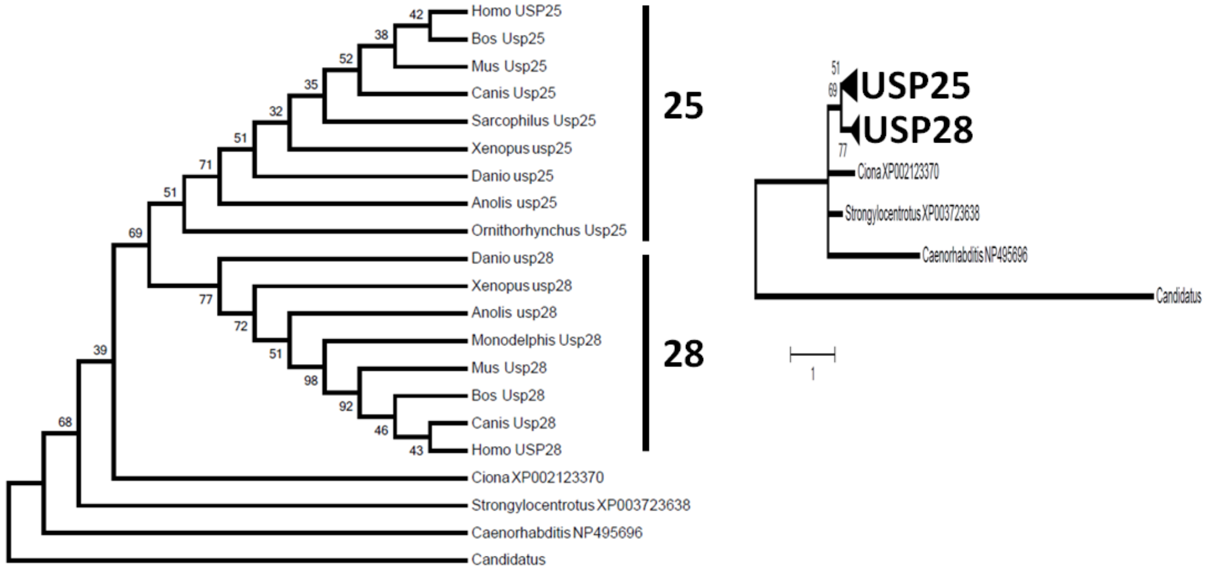


Figure 5.7. Phylogenetic history of group 10 USP homologues. (A) A phylogenetic tree was reconstructed by the maximum likelihood method using C19 domain protein sequences, adopting the Jones Taylor and Thornton model with gamma distribution and with some invariable sites (JTT+G+I) (Jones *et al.*, 1992). All vertebrate homologues are annotated according to their orthologous relationship as found in studies, whereas NV homologues are represented by the accession numbers. Branch length format of the same is show in **(B)**.

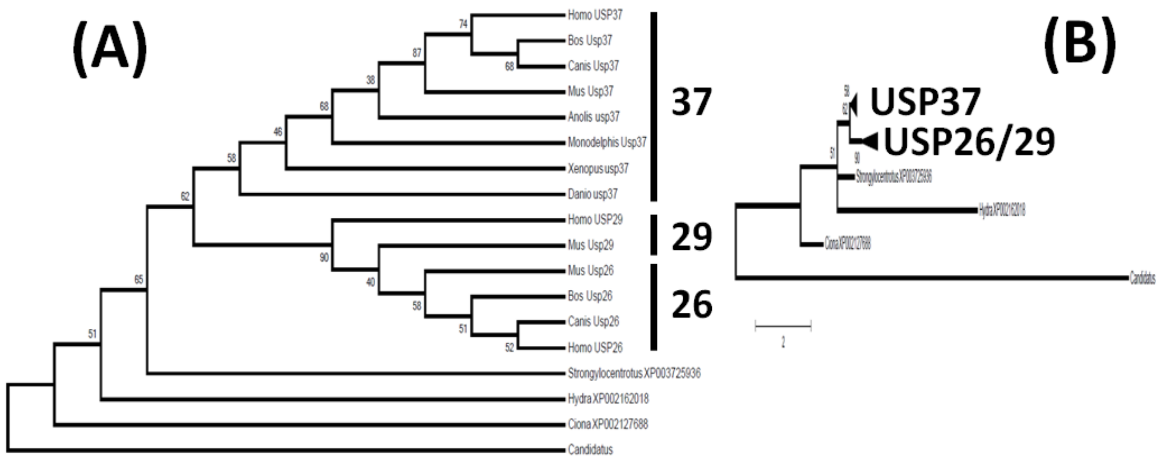


Figure 5.8. Phylogenetic history of group 11 USP homologues. (A) A phylogenetic tree was reconstructed by the maximum likelihood method using C19 domain protein sequences, adopting the Jones Taylor and Thornton model with gamma distribution and with some invariable sites (JTT+G+I) (Jones *et al*., 1992). All vertebrate homologues are annotated according to their orthologous relationship, whereas NV homologues are represented by the accession numbers. Branch length format of the same tree is shown in **(B)**.

Group 15: Group 15 includes: USP53 and USP54. Vertebrate USP53 and USP54 form separate subclades in phylogenetic tree reconstruction and share a lineal ancestry with one homologue of tunicate, sea urchin and fruit fly (Figure 5.9) pointing to the separation of USP53 and USP54 at the root vertebrates from a single ancestral NV gene.

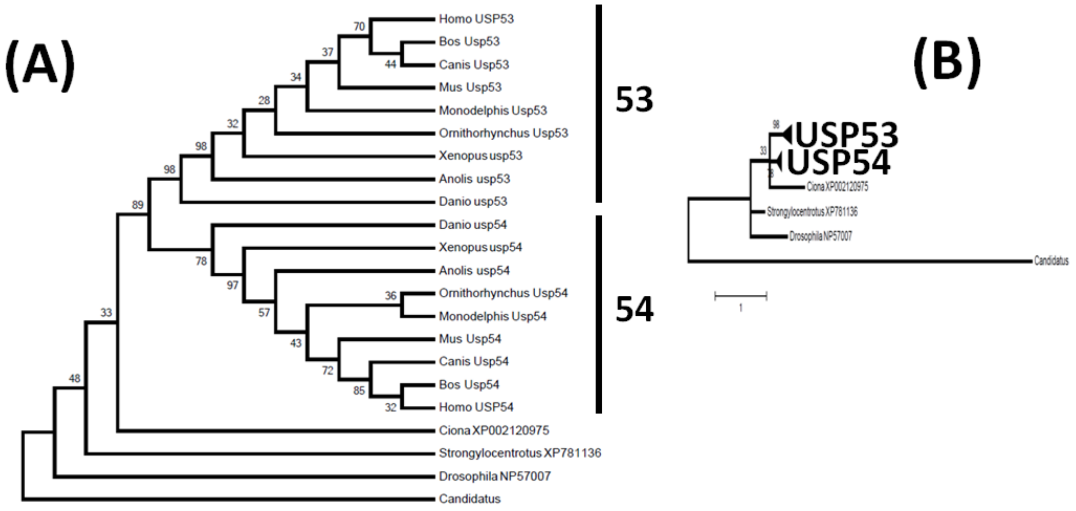


Figure 5.9. Phylogenetic history of group 15 USP homologues. (A) A phylogenetic tree was reconstructed by the maximum likelihood method using C19 domain protein sequences, adopting the Jones Taylor and Thornton model with gamma distribution and with some invariable sites (JTT+G+I) (Jones et al., 1992). All vertebrate homologues are annotated according to their orthologous relationship, whereas NV homologues are represented by the accession numbers. Branch length format of the same tree is shown in **(B)**.

5.5. Genomic synteny reflects multiple mechanisms underlying expansion of USPs

The extensive expansion of USPs in early vertebrates probably reflects the two whole genome duplication (WGD) events at the base of vertebrates. The resulting homologues are referred to as ohnologues (Wolfe, 2000). However, other mechanisms such as gene duplication, segmental duplication within chromosomes and gene domain acquisition may also have contributed to the expansion and diversification of these genes. To explore these possibilities, adjacent genomic regions of the human USP loci were compared (examining 10 genes on either side) (Figure 5.10).

Genome Duplication. Datamining and phylogenetic analyses presented in this study reflect the expansion of USPs at the origin of vertebrates, indicative of the expansion of USPs via WGD. Since WGD events generated duplicate copies of chromosomes harbouring identical copies of genes (at least initially), it is reasoned that many of the duplicates underwent gene death, mainly because of redundancy or neutral drift (Innan and Kordrashov, 2010). However, the genes retained (possibly due to functional divergence) in the organisms, despite being present on different chromosomes remains phylogenetically linked. Moreover, if not being subjected to genomic rearrangement, several of these genes have genes neighbours showing a syntenic relationship (having been duplicated by the same event). Several phylogenetically linked USP loci in the vertebrate lineage are associated with syntenic paralogues, for example in humans, USP12 and USP46 are present on chromosome 13q and 4q respectively and are neighboured by genes from three other evolutionary related families namely: Small Nucleolar RNA (SNORA), Ras like (RASL) and Ligand of Numb protein X (LNX). This situation result from the consequence of WGD. Similarly, USP35 (chromosome 11q) and USP38 (chromosome 4q) are adjacent to genes encoding GRB2 associated Binding Proteins (GAB). USP2 and USP21 are present on chromosomes 1q and 15q respectively and both loci contain one homologue of an evolutionary related gene family Polio Virus Receptor Like protein (PVRL). Human USP36 and USP42 are present on chromosomes 17q and 7p respectively and adjacent to cytohesin encoding homologues. USP33 and USP20, present on chromosomes 1p and 9q respectively, are neighboured by Far Upstream Binding Protein (FUBP) homologues. Both members of group 5, USP5 and USP13, are present on different chromosomes (12p and 3q respectively) and flanked by Guanine Nucleotide Binding protein (GNB) homologues. Finally, in humans, USP53 and USP54 are present on chromosomes 4q and 10q respectively and are located downstream of three evolutionary related gene families, namely myozenin MYOZ), synaptopodin (SYNOP) and SEC24 family genes (SEC24).

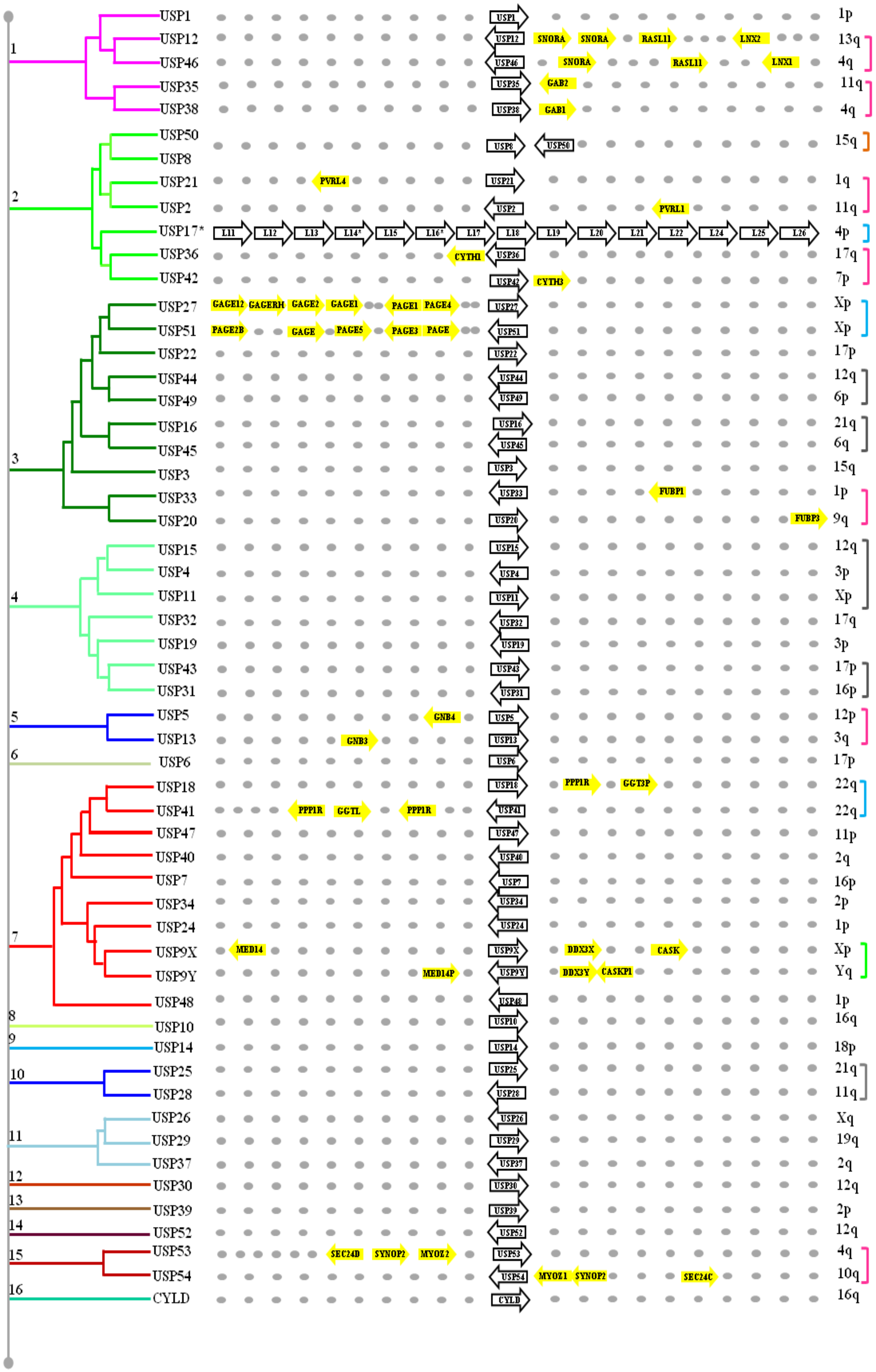

1
USP1
USP12
USP46
USP35
USP38
2
USP50
USP8
USP21
USP2
USP17*
USP36
USP42
3
USP27
USP51
USP22
USP44
USP49
USP16
USP45
USP3
USP33
USP20
4
USP15
USP4
USP11
USP32
USP19
USP43
USP31
5
USP5
USP13
6
USP6
7
USP18
USP41
USP47
USP40
USP7
USP34
USP24
USP9X
USP9Y
USP48
8
USP10
9
USP14
10
USP25
USP28
11
USP26
USP29
USP37
12
USP30
13
USP39
14
USP52
15
USP53
USP54
16
CYLD
SNORA
SNORA
RASL11
LNX2
SNORA
RASL11
LNX1
GAB2
GAB1
PVRL4
PVRL1
L11
L12
L13
L14*
L15
L16*
L17
L18
L19
L20
L21
L22
L24
L25
L26
CYTH1
CYTH3
GAGE12
GAGERH
GAGE2
GAGE1
PAGE1
PAGE4
PAGE2B
GAGE
PAGE5
PAGE3
PAGE
FUBP1
FUBP3
GNB4
GNB3
PPP1R
GGT3P
PPP1R
GGTL
PPP1R
MED14
DDX3X
CASK
MED14P
DDX3Y
CASKP1
SEC24D
SYNOP2
MYOZ2
MYOZ1
SYNOP2
SEC24C
1p
13q
4q
11q
4q
15q
1q
11q
4p
17q
7p
Xp
Xp
17p
12q
6p
21q
6q
15q
1p
9q
12q
3p
Xp
17q
3p
17p
16p
12p
3q
17p
22q
22q
11p
2q
16p
2p
1p
Xp
Yq
1p
16q
18p
21q
11q
Xq
19q
2q
12q
2p
12q
4q
10q
16q

Figure 5.10. Genomic Synteny of Human USPs homologues: The genomic context of the human USP homologues is shown. The inferred evolutionary relationship is indicated at the left hand side with their respective paralogous group number. Syntenic genes are highlighted as yellow arrows while the USP homologues are represented with open arrows. The arrow direction corresponds to the direction of transcription with respect to the chromosome map. The gray dots indicate irrelevant or non syntenic genes to either side of the USP. The possible mechanism of genes expansion is represented by colour coded parenthesis to the right pink: potential WGD event (where syntenic evidence is present); gray: potential WGD events (where no syntenic evidence was found); orange: gene duplication; blue: segmental duplication; green: chromosomal conversion. Note, all genes of USP17* are not shown.

Other USP paralogous pairs: USP44/USP49, USP16/USP45, USP43/USP31, USP25/USP28 and a triplet USP4, USP11&USP15 showed a divergence in early vertebrates (according to the trees) and are present on different chromosomes. However, no syntenic paralogous gene (within 10 adjacent genes) was identified associated with these paralogous USPs. The absence of associated paralogous genomic synteny may reflect the extensive genomic rearrangement postdating WGD, or gene loss of the linked gene.

Segmental duplication. Two gene pairs, USP27/USP51 and USP18/USP41 are phylogenetically linked, present on the same chromosomes and share syntenic regions (Figure 5.10). The USP27 and USP51 pair are specific to eutherian mammals and are present at different regions of same chromosome (Xp). Each gene is proximal to P Antigen family Genes (PAGE) and G Antigen family Genes (GAGE). Similarly, USP18 and USP41 are present at different regions of chromosome 22q and are proximal to Protein Phosphatase 1 Regulatory unit Pseudogenes (PPP1RP) and Gamma Glutamyltransferase Pseudogenes (GGT3P). The synteny shown by these USP pairs on the same chromosome strongly suggest extrachromosomal segmental duplication as the mechanism of expansion.

Another interesting example in this connection is the presence of 32 copies of USP17 genes and pseudogenes which are located on chromosomes 4p (23 copies) and 8p (9 copies) in humans. The same scenario of multiple copies of USP17 gene exists in other eutherian mammals, including the mouse in which at least six intact orf and one pseudogenes of USP17 exists. Earlier phylogenetic analysis have shown that the paralogous based clustering in the USP17 gene family points to species specific extensive gene duplications and/or tandem segmental duplications, which lead to the formation of multiple copies of USP17 (Burrows *et al.*, 2010).

Gene duplication. USP50 has a vertebrate specific distribution and phylogenetically it is closely related to USP8. In humans the two genes are located head to head on chromosome

15q suggesting to the origin of USP50 from a gene duplication of the ancestral gene, most likely similar to USP8 in early vertebrates.

Chromosomal conversion. It is now widely accepted that the X and Y chromosome were the homologous autonomies and were converted into sex chromosomes after the monotremes-marsupial split (Grave *et al*., 2006; Lehn and Page, 1999). USP9X and USP9Y are phylogenetically related and present on Xp and Ya chromosomes respectively. Loci of the two genes are flanked by paralogues that include: DEAD box helicase (DDX3X), Calcium/Calmodulin dependent Serine protein Kinase (CASK) and MED14 mediator complex subunit 14 (MED14) genes. This suggests the USP9Y and its proximal syntenic regions are among the few genes that have been retained in the Y chromosome since its existence.

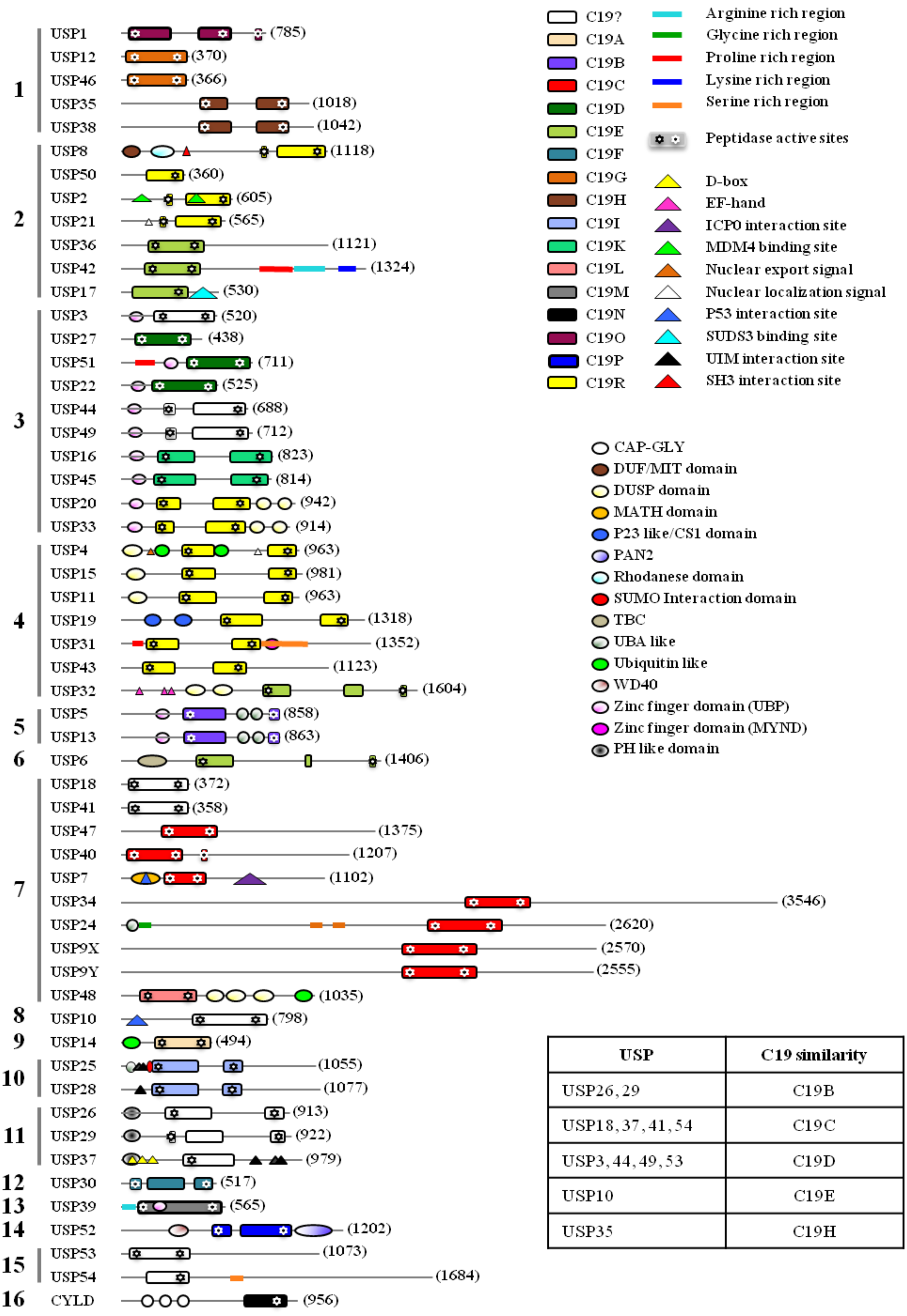
USP26 and USP29 are specific to eutherian mammals and shows close relationship with the USP37. However, the genomic synteny does not provide any evidence of gene or segmental duplication in relation to their origin. It is possible that these genes may have arisen due to the domain shuffling of USP37 or indeed by gene duplication and subsequently subjected to genetic rearrangement.

Taken together, the data suggest that the genome duplication events contribute substantially to the expansion of USPs extant in humans today. In addition to this other mechanisms such as gene and intra chromosomal segmental duplications also lead to the origin of certain USPs. However, domain acquisition combined with serendipitous genetic rearrangement may provide alternative and perhaps counterintuitive explanation for the expansion of USPs especially at the root vertebrate.

5.6. Protein domain promiscuity in USPs

A protein domain can be broadly defined as a sequentially and/or structurally conserved region that can evolve, function or exist independently from the rest of the protein sequence. Protein domains occur either as a single entity within a coding sequence or in combinations and different permutations with the other domains, a feature referred to as “domain versatility” or “promiscuity” (Basu *et al*., 2008). In order to examine the protein structural and/or functional domain diversity of USPs, the protein sequences were assessed using the conserved domain database (CDD) and UniProt database (Figure 5.11 and Table 5.3). For clarity, domain variations are described separately for each paralogous group.

Group 1. Among the paralogues of group 1, USP12 and USP46 both have a C19G domain, supporting their close phylogenetic relationship, while USP38 and USP35 contain C19H. The C19 domains of USP35 and USP38 are both split into two, also reflecting their close phylogenetic relationship. Consistent with the phylogenetic studies, USP1 bears a C19O



USP	C19 similarity
USP26, 29	C19B
USP18, 37, 41, 54	C19C
USP3, 44, 49, 53	C19D
USP10	C19E
USP35	C19H

Figure 5.11. Protein domain distributions in human USPs. Schematic depiction of protein domains of the human USPs is shown. The length of each homologue is scaled according to the size of the protein and total amino acid length is indicated in bracket. The paralogous group number is indicated on the left. A key for the different protein domains and motifs is given on the right top. The table inside the figure shows the similarity between the un typed C19 domains with the subtyped C19 domains.

domain which is split into three in human orthologue.

Group 2. Two types of C19 domain are present in group 2 paralogues: C19R and C19E. The C19R domain (split into two except in USP50) was found in USP8, USP50, USP2 and USP21 while intact C19E was observed in USP17, USP36 and USP42. This C19 domain variation between the paralogues is in agreement with their phylogenetic relationship, as members with C19R and C19E are separated in the first dichotomy (excluding outgroup) of the tree. In addition to the C19 domain, other domains and functional sites are present in different paralogues of group 2. For example, USP8 contains a rhodanese domain and a domain of unknown function (DUF) which are absent in its phylogenetically closest paralogue, USP50. As described above, USP50 originated by gene duplication suggesting the additional domain and part of C19R were lost from the USP50 paralogue. Similarly, USP2, has an MDM4 binding site, absent from USP21 which instead has a nuclear localization signal. Proline, arginine and lysine rich sequences are present in USP42 which are not found in the closest paralogue USP36, while USP17 (closely related to 36/42) instead has a SUDS3 binding domain.

Group 3. Group 3 is the most diverse group with respect to the variations in the C19 domain. Consistent with the phylogenetic analysis, USP27, USP51 and USP22 all contain C19D domain while USP16 and USP45 have a split C19K domain. Similarly, USP33 and USP20 (phylogenetically more close than any other member of this group) contain a split C19R domain. Three paralogues of this group, USP44, USP49 and USP3 have un subtyped C19. Among those, USP44 and USP49 C19 domain asymmetrically split into two. By comparison, the C19 domain of USP3 is intact and phylogenetically it forms a distinct monophyletic clade suggesting that this is a different variant of the C19 domain. C19 domains of these homologues (USP3, USP44 and USP49) show similarity with C19D type domain. USP33 and USP20 (C19R) have two additional DUSP domains, this is in line with their close phylogenetic relationship. Interestingly, all paralogues of group 3 except USP27 (which is eutherian specific) contain a zinc finger domain, suggesting that this domain existed in the ancestral gene of group 3 before expansion.

Group 4. All paralogues of group 4, except USP32, have a C19R domain which is split into two. The C19 domain of USP32 is split into three and is not subtyped in CDD. In addition, USP32 contains three calcium binding (EF-hand) sites and a DUSP domain. USP4, USP11 and USP15, which may have expanded from a single lineal ancestral gene at the base of vertebrates, all contain a DUSP domain, suggesting the presence of the DUSP domain in the ancestral gene of all three paralogues. Additionally, USP4 contains two ubiquitin like domains, two zinc finger domains and nuclear localization and export sequences, which possibly reflect the domain and functional sites acquisition after gene expansion. USP19, which forms a separate monophyletic clade in the phylogenetic tree, also contains a zinc finger domain and two p23 like domains. The phylogenetically close paralogues, USP43 and USP31, have no additional domain except a stretch of proline and serine rich sequences in USP31.

The presence of a C19R domain in some members of groups 2, 3 and 4 may be an indicative of a close evolutionary relationship between these groups than the other groups of USPs. Given the homologues of all these groups were found in slime mould, suggest that this division occurred before protozoa and metazoa split.

Group 5. Both paralogues of group 5, USP5 and USP13, contain C19B domain (exclusive to this group) which is split into two and intervened by 2 UBA like domains. The paralogues also contain a zinc finger domain and the ORF size and domain structure are highly similar between the two genes supporting their close phylogenetic association.

Group 6. USP6 is the sole member of group 6 which contains a C19 domain, which is split into three. Other non USP members of this group contain a TBC domain, which may form the basis of paralogous relationship between USP6 and other TBC domain containing proteins. As USP6 homologues are only present in primates, it suggests a relatively recent gene birth, with the acquisition of a C19 domain in the primate lineage.

The C19 domain structure of USP6 bears a remarkable resemblance to that of USP32. Therefore in order to explore their relatedness, USP6 was included in the phylogenetic tree of group4. The USP6 C19 domain clusters in the USP32 clade (Figure 5.12A). Thus it seems likely that USP6 arose in a primate lineage (leading to the greater apes) through the acquisition of a USP32 C19 domain into a TBC encoding gene.

Group 7. Paralogues of group 7 encode some of the largest in terms of protein length. Most of them contain an intact C19C domain (split only USP40), the only exception in this regard is USP48 and USP18/USP41 which contain C19L and un subtyped C19 domains respectively and are more distantly related to others in the group, however, C19 domain of USP18/41 shows similarity with C19C. USP7 is unique among all USPs in containing a MATH domain and a herpes virus transactivator, ICP0 binding site. MATH domain interacts with the EBV oncogenic protein EBNA and the tumour suppressor protein, p53. The phylogenetically related USP40 and USP47 contain none of these additional domains.

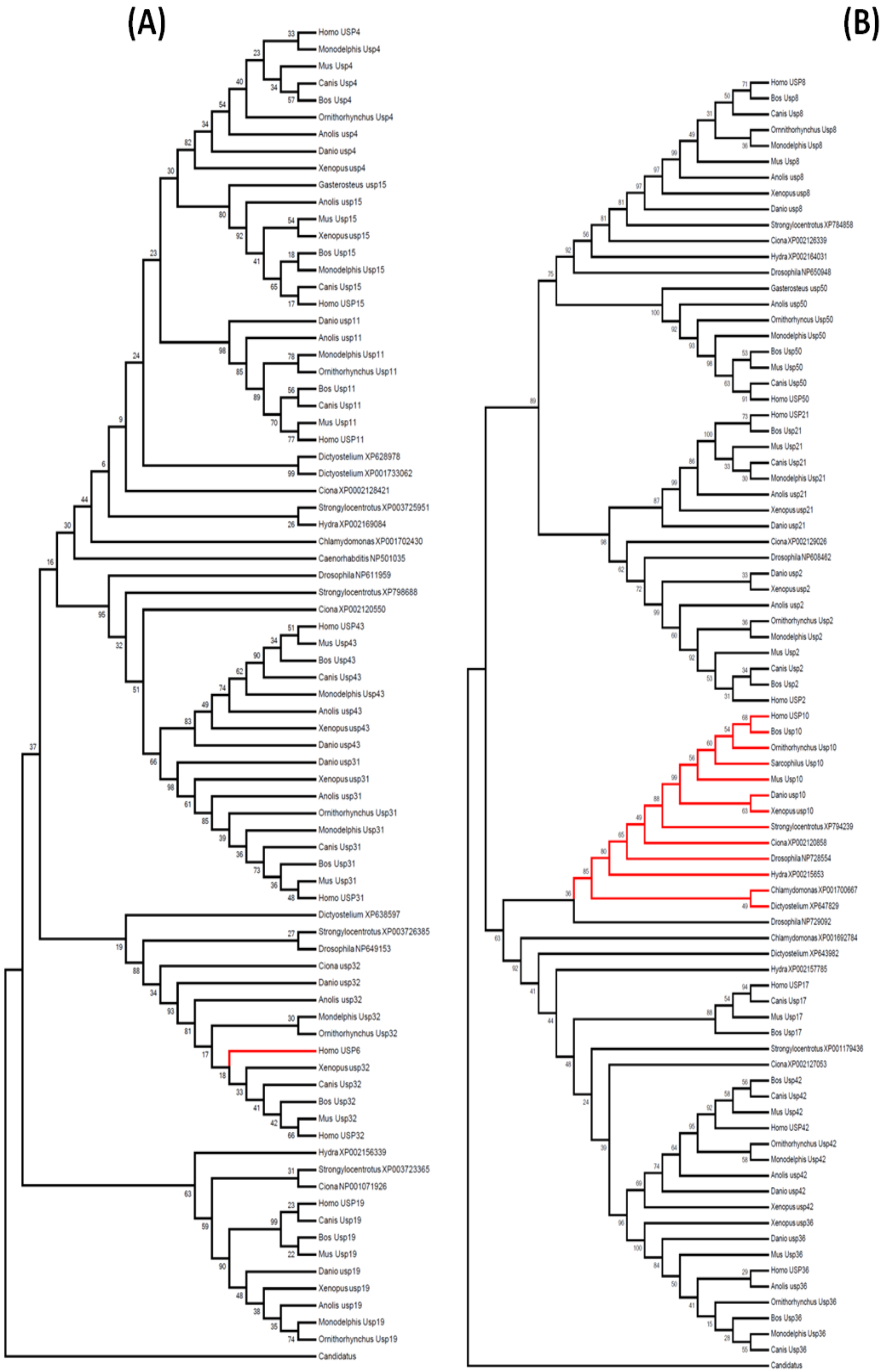

(A)
Homo USP4
Monodelphis Usp4
Mus Usp4
Canis Usp4
Bos Usp4
Ornithorhynchus Usp4
Anolis usp4
Danio usp4
Xenopus usp4
Gasterosteus usp15
Anolis usp15
Mus Usp15
Xenopus usp15
Bos Usp15
Monodelphis Usp15
Canis Usp15
Homo USP15
Danio usp11
Anolis usp11
Monodelphis Usp11
Ornithorhynchus Usp11
Bos Usp11
Canis Usp11
Mus Usp11
Homo USP11
Dictyostelium XP628978
Dictyostelium XP001733062
Ciona XP0002128421
Strongylocentrotus XP003725951
Hydra XP002169084
Chlamydomonas XP001702430
Caenorhabditis NP501035
Drosophila NP611959
Strongylocentrotus XP798688
Ciona XP002120550
Homo USP43
Mus Usp43
Bos Usp43
Canis Usp43
Monodelphis Usp43
Anolis usp43
Xenopus usp43
Danio usp43
Danio usp31
Xenopus usp31
Anolis usp31
Ornithorhynchus Usp31
Monodelphis Usp31
Canis Usp31
Bos Usp31
Mus Usp31
Homo USP31
Dictyostelium XP638597
Strongylocentrotus XP003726385
Drosophila NP649153
Ciona usp32
Danio usp32
Anolis usp32
Mondelphis Usp32
Ornithorhynchus Usp32
Homo USP6
Xenopus usp32
Canis Usp32
Bos Usp32
Mus Usp32
Homo USP32
Hydra XP002156339
Strongylocentrotus XP003723365
Ciona NP001071926
Homo USP19
Canis Usp19
Bos Usp19
Mus Usp19
Danio usp19
Xenopus usp19
Anolis usp19
Monodelphis Usp19
Ornithorhynchus Usp19
Candidatus
(B)
Homo USP8
Bos Usp8
Canis Usp8
Ornithorhynchus Usp8
Monodelphis Usp8
Mus Usp8
Anolis usp8
Xenopus usp8
Danio usp8
Strongylocentrotus XP784858
Ciona XP002126339
Hydra XP002164031
Drosophila NP650948
Gasterosteus usp50
Anolis usp50
Ornithorhyncus Usp50
Monodelphis Usp50
Bos Usp50
Mus Usp50
Canis Usp50
Homo USP50
Homo USP21
Bos Usp21
Mus Usp21
Canis Usp21
Monodelphis Usp21
Anolis usp21
Xenopus usp21
Danio usp21
Ciona XP002129026
Drosophila NP608462
Danio usp2
Xenopus usp2
Anolis usp2
Ornithorhynchus Usp2
Monodelphis Usp2
Mus Usp2
Canis Usp2
Bos Usp2
Homo USP2
Homo USP10
Bos Usp10
Ornithorhynchus Usp10
Sarcophilus Usp10
Mus Usp10
Danio usp10
Xenopus usp10
Strongylocentrotus XP794239
Ciona XP002120858
Drosophila NP728554
Hydra XP00215653
Chlamydomonas XP001700667
Dictyostelium XP647829
Drosophila NP729092
Chlamydomonas XP001692784
Dictyostelium XP643982
Hydra XP002157785
Homo USP17
Canis Usp17
Mus Usp17
Bos Usp17
Strongylocentrotus XP001179436
Ciona XP002127053
Bos Usp42
Canis Usp42
Mus Usp42
Homo USP42
Ornithorhynchus Usp42
Monodelphis Usp42
Anolis usp42
Danio usp42
Xenopus usp42
Xenopus usp36
Danio usp36
Mus Usp36
Homo USP36
Anolis usp36
Ornithorhynchus Usp36
Bos Usp36
Monodelphis Usp36
Canis Usp36
Candidatus

Figure 5.12. Phylogenetic relationship between paralogous groups of USPs. The phylogenetic trees were reconstructed using the maximum likelihood method and employing JTT+G **(A)** or JTT+G+I **(B)** evolutionary model. The phylogenetic trees show the relationship between paralogous group 4 and group 6 **(A)** and group 2 and group 8 **(B)**. The orthologues of group 6 and group 8 are indicated with red lines while homologues of group 2 and group 4 are coloured black.

none of these additional domains. USP34 is the largest among all USPs examined but interestingly contains only C19C domain. Other uncharacterized or unidentified functional sites or domains could be present in this and other USPs particularly USP24, USP9X and USPY which are also large, however USP24 does have UBA domain in it. The USP48 has three DUSP domains and one DUF domain and formed a separate clade in the phylogenetic analysis.

Group 8. The C19 domain of USP10, the only member of group 8, is intact and unsubtyped. The protein contains a binding site for p53 protein towards the N-terminus. The domain show similarity with C19E subtype, a domain also found among members of group 2. To examine it further, a composite tree was developed using homologues of group 2 and group 8 (Figure 5.11B). In this phylogenetic reconstruction, this group cluster with USP36/45 suggesting the close evolutionary relationship between group 2 and group 8.

Group 9. The only member of this group is USP14, which exclusively contains a C19A domain and a DUF domain.

Group 10. USP25 and USP28 are the two paralogues comprising the group 10, both containing C19I domains which are split into two and both also contain a UIM interaction site, supporting their close phylogenetic relationship. However, USP25 has an additional DUSP domain and SUMO interaction site reflecting acquisition or loss of new functional sites in the gene after duplication.

Group 11. USP26, USP29 and USP37, the members of group 11, each contain an un subtyped C19 domain which is either intact (USP37) or split into two (USP26) or three (USP29) regions. USP37 also contains three destruction box domain (D-box), and UIM interaction sites. Given the species distribution of USP26 and USP29, limited to eutherian mammals, while orthologues of USP37 were observed in all vertebrates, this suggests the loss of some functional sites in USP26 and USP29 after expansion of ancestral USP37 gene.

Group 12. USP30 is the single member of group 12 containing a C19F domain (exclusive to this group) which is split into three segments.

Group 13. Group 13 contains only one member, USP39, which has peptidase C19M domain with an additional DUSP domain within the C19 sequences and an arginine rich region at the N-terminus.

Group 14. USP52 is present in all vertebrates and a nonvertebrate examined and contains a C19P domain, which is split into two regions. In addition it contains PAN2 and WD40 domain sequences. These domains are unique to USP52 among the human USPs.

Group 15. Two paralogues of group 15, USP53 and USP54 contain an un subtyped C19 domain without any additional domain. The two proteins differ in their length which may reflect the structural and/or functional differences between these proteins, and the C19 domain of USP54 has lost the N-terminal peptidase active site.

Group 16. CYLD is the only member of this group, contains exclusively a C19N domain with three CAP-GLY domains.

In summary, the C19 domain and domain structure of the USPs show broad agreement with the phylogenetic analysis, where closely related paralogues have the same C19 domain subtype. The variation in the additional domains (other than C19) between the paralogues reflects the functional divergence between the USPs.

It has been proposed that protein architectural complexity is correlated with organismal complexity (Koonin *et al*., 2002; Vogel *et al*., 2004). To explore this proposition and

to investigate the evolutionary history of the domain distribution, domain combinations of USPs were examined in selected species using InterPro, CDD and UniProt databases. The domain distribution is tabulated according to the species in which they were identified (Table 5.3). USP homologues of slime mould and chlamydomonas were investigated for

their domain distribution. Chlamydomonas (11 USPs) and slime mould (17 USPs) are unicellular organisms and serve as models for the ancestral lineages of plants and animals respectively. Of the 16 C19 subtypes (excluding unclassified) in humans, 9 are present in both chlamydomonas and slime mould, namely: C19G (USP12&USP46 of group 1), C19E (USP17, USP36 and USP42 of group 2), C19R (group 2, 3 and 4), C19B (group 5), C19C (group 7), C19L (USP48 of group 7) C19A (group 9), C19M (group 13) and C19P (group 14). This observation is consistent with the datamining and phylogenomic analysis suggesting the origin of most USP paralogous groups before the divergence of animals and plants. In addition, combinations of domains: C19A+DUF (USP14), C19C+MATH (USP7), C19B+Zf+UBA (USP5, USP13) and C19P+PAN2 (USP52) were observed amongst the USP homologues of both slime mould and chlamydomonas, supporting the orthologous relationship observed in the C19 domain based phylogenetic analysis between these genes and their corresponding counterparts in vertebrates including humans. Unlike chlamydomonas, 2 additional C19 domains were observed in slime mould; C19H (USP35 & USP38 of group 1) and C19 K (USP16&USP45 of group 3). Additionally, 4 further domain combinations were found namely: WD40+C19P+PAN2 (USP52), USP19R+DUSP (group3,4), C19C+MATH+ICP0 (USP7), and C19L+DUSPx3 (USP48) in the USP

C19	Cr	Dd	Hm	Ce/Dm	Sp	Ci	Vertebrates	Eutheria	Primates
C19A	DUF	DUF	DUF	DUF	DUF	DUF	DUF	DUF	DUF
C19B	UBAx1+Zf	UBAx2+Zf	UBAx2+Zf	UBAx2+Zf	UBAx2+Zf	UBAx2+Zf	UBAx2+Zf	UBAx2+Zf	UBAx2+Zf
C19C	MATH	MATH+ICP0	MATH+ICP0 UBA	MATH+ICP0 UBA	MATH+ICP0 UBA	MATH+ICP0	MATH+ICP0 UBA	MATH+ICP0 UBA	MATH+ICP0 UBA
C19D			Zf	Zf	Zf	Zf	Zf	Zf	Zf
C19E	🟨	🟨	🟨	🟨	🟨	🟨	🟨	SUDS3	SUDS3
C19F				🟨	🟨	🟨	🟨	🟨	🟨
C19G	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
C19H		🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
C19I				🟨	🟨	🟨	UBA	UBA	UBA
C19K		🟨	🟨	Zf	Zf	🟨	Zf	Zf	Zf
C19L	Cyclase	DUSPx3	?	?	DUSPx3+UBA	DUSPx3+UBA	DUSPx3+UBA	DUSPx3+UBA	DUSPx3+UBA
C19M	Zf	Zf	Zf	Zf	Zf	Zf	Zf	Zf	Zf
C19N				CAP-GLY	CAP-GLYx3	CAP-GLYx3	CAP-GLYx3	CAP-GLYx3	CAP-GLYx3
C19O				🟨	🟨	🟨	🟨	🟨	🟨
C19P	PAN2	PAN2+WD40	PAN2+WD40	PAN2+WD40	PAN2+WD40	PAN2+WD40	PAN2+WD40	PAN2+WD40	PAN2+WD40
C19R	🟨	DUSP	?	DUSP DUSPx2 DUF+Rhodanese EF-hand	DUSP DUSPx2 DUF+Rhodanese EF-hand P23like	DUSP DUSPx2 DUF+Rhodanese EF-hand P23like	DUSP Zf+DUSPx2 DUF+Rhodanese EF-hand P23like MDM4 DUSP+UbX2	DUSP Zf+DUSPx2 DUF+Rhodanese EF-hand P23like MDM4 DUSP+UbX2	DUSP Zf+DUSPx2 DUF+Rhodanese EF-hand P23like MDM4 DUSP+UbX2
C19? (44/49)						Zf	Zf	Zf	Zf
C19?(6)									TBC
C19?(18/41)							🟨	🟨	🟨
C19?(10)	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
C19? (26/27/37)			🟨		DBOX UIM interaction	DBOX UIM interaction	DBOX UIM interaction	DBOX UIM interaction	DBOX UIM interaction
C19? (53/54)				🟨	🟨	🟨	🟨	🟨	🟨

Table 5.3. Origin and distribution of domains and domains combinations associated with USPs. InterPro, CDD and UniProt databases were employed to screen the domains present in USP homologues of selected species. The yellow boxes indicate the domain identified in the respective orthologues of USPs. Different combinations of domains are indicated in the boxes. The abbreviations of the domains are indicated in figure 5.11.

homologues of slime mould, suggesting a premetazoan origin of most USP paralogous groups.

In the animal lineage the first major transition happened with the emergence of metazoans around 1184MYA. In this analysis two domain combinations (including the new C19D variant) were observed in the USP homologues of hydra. These are: C19C+UBA and C19D+Zf, found in USP24 and USP22, respectively. Both these additional domains (other than C19) pre-exist in different USP homologues in slime mould. USP homologues in the fruit fly, in addition to the already existing domain combinations (as found in representatives of early branches of the tree), have 3 new variants of C19, C19O (USP1), C19F (USP30) and C19N (CYLD), and 5 new domain combinations were found which persist in higher organisms, C19R+DUF+Rhodanese (USP8), C19N+CAP-GLY (CYLD), C19R+DUSPx2 (USP33/20), C19R+DUF+EF-hand (USP3), C19K+Zf (USP16/45). This suggests an increase in the domain permutations occurred with the origin of coelomates. The sea urchin genome encodes 31 USP homologues, of which three new domain arrangements were observed: p23+C19R (USP19), DUSPx3+UBA+C19L (USP48) and CAP-GLYx3+C19N. These combinations are also found in higher organisms. Additionally, a domain combination of C19?+DBOX+UIM interaction was also first observed in the sea urchin homologue of contains three calcium binding (EF-hand) sites and DUSP domain. Among the additional domains p23, DBOX and the UIM interaction site, were not identified in any USP homologue of organisms whose lineage emerged before the divergence of deuterostomes, indicating a further increase in the domain complexity as organismal complexity increases from protostomes to deuterostomes. However, the increment in the CAP-GLY (CYLD) is most likely the result of internal domain duplication.

In chordates (specifically the tunicate examined here) no significant increase in the domain combinations was observed compared to earlier organisms, the only novelty observed is an emergence of a variant (un subtyped) of C19 domain (USP44/49) with zinc finger. However, unsubtyped C19 domain of USP44/49 show the greatest similarity to C19D (found also in USPs of group 3). New domain arrangements that emerged with the origin of vertebrates include: C19I+UBA (USP25), C19R+MDM4 (USP2) C19R+DUSP+Ubx2 (USP4) and C19R+DUSPx2+Zf (USP20&USP33). Further, two new domain combinations, C19D+SUDS3 (USP17) and C19+TBC (USP6) appeared with the origin of eutherian mammals and primates respectively.

Taken together, similarities in domain arrangements between non vertebrates and the vertebrate USPs supports their phylogenetic relationship. C19 domain variations and domain combinations in USPs increase across the animal lineages especially in species reflecting the important transitional stages of speciation. This mirrors the increase in molecular complexity from early unicellular eukaryotes to primates.

5.7. Structural similarities of peptidase C19 domain

In order to compare the structure of the different peptidase C19 subtypes, the structures of several USPs were retrieved from the RCSB protein databank. Primary and tertiary structure alignments were developed and quantified in terms of sequence identity in percentage and root mean square deviation (RMSD) of the structures in Å (Table 5.4; 5.5; Figure 5.14). Despite limited primary sequence identity (6%-36%) between the peptidase C19 domains analysed (Table 5.4), Cα backbone superimposition of the domains shows considerable similarity between the structures in the core architecture (RMSD values range from 1.02Å to 2.09Å, Table 5.5). Structurally, the C19 domain is composed of three subdomains termed the palm, thumb and fingers (Hu *et al.*, 2002). The C19 domain structures have the same characteristic sub-domains with nearly identical distribution of the secondary structural elements. The only exception is the C19 domain of CYLD in which all subdomains are shorter than those of the other paralogues. This strong conservation of the domain and subdomain architecture reflects the common ancestral origin of the C19 domain as well as conserved enzymatic action (Figure 5.14).

	USP2 (C19R)	USP4 (C19R)	USP5 (C19B)	USP7 (C19C)	USP14 (C19A)
USP4	40%				
USP5	16%	16%			
USP7	20%	18%	13%		
USP14	17%	18%	14%	18%	
CYLD	9%	8%	9%	9%	7%

Table 5.4. Sequence similarity between the selected C19 domains. Multiple amino acid sequence alignment of the peptidase C19 domains of the indicated human USP paralogues (of which structures are known) were constructed using ClustalX and the amino acid identities between these are shown.

	USP2 (C19R)	USP4 (C19R)	USP5 (C19B)	USP7 (C19C)	USP14 (C19A)
USP4	1.02				
USP5	1.10	1.32			
USP7	1.31	1.36	1.27		
USP14	1.28	1.20	1.27	1.35	
CYLD	2.09	1.64	2.09	1.64	1.66

Table 5.5. Structural comparison between selected C19 domains. Protein structures of the indicated human USPs were retrieved from the RCSB database, only the C19 domains were extracted from the atomic coordinates. Structures were superimposed and variations in Cα back bone were measured in terms of RMSD values in Å.

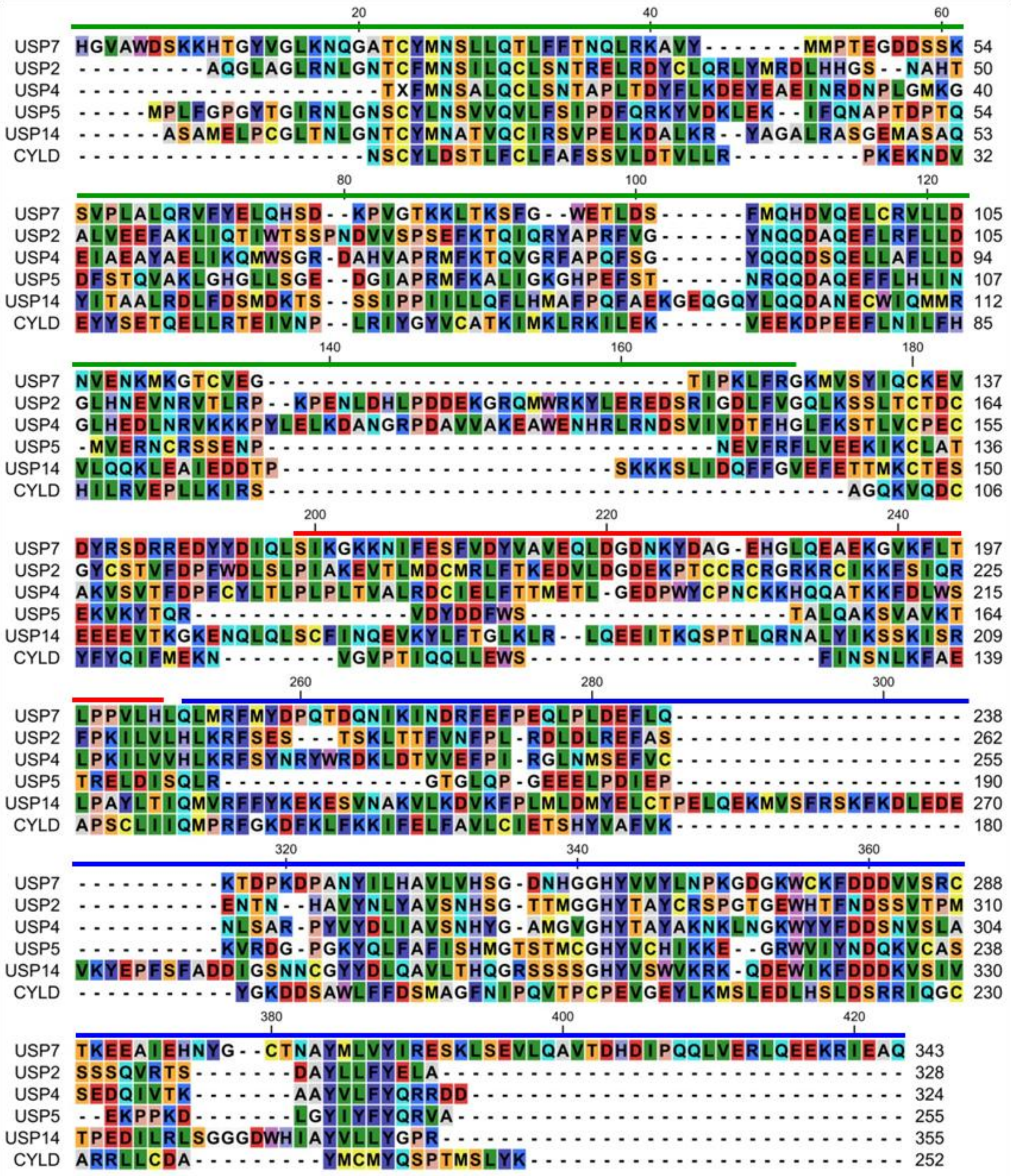


Figure 5.13. Multiple sequence alignment of C19 domain. Primary sequences of C19 domain (of those which are structured) are aligned using CLUSTALX. the amino acids are colour coded according to RASMOL convention. The bars over the top of the alignment represent different structural regions as observed in the USP7 C19 domain (PDBid; 1F1Z): green (thumb), red (fingers) and blue (palm).

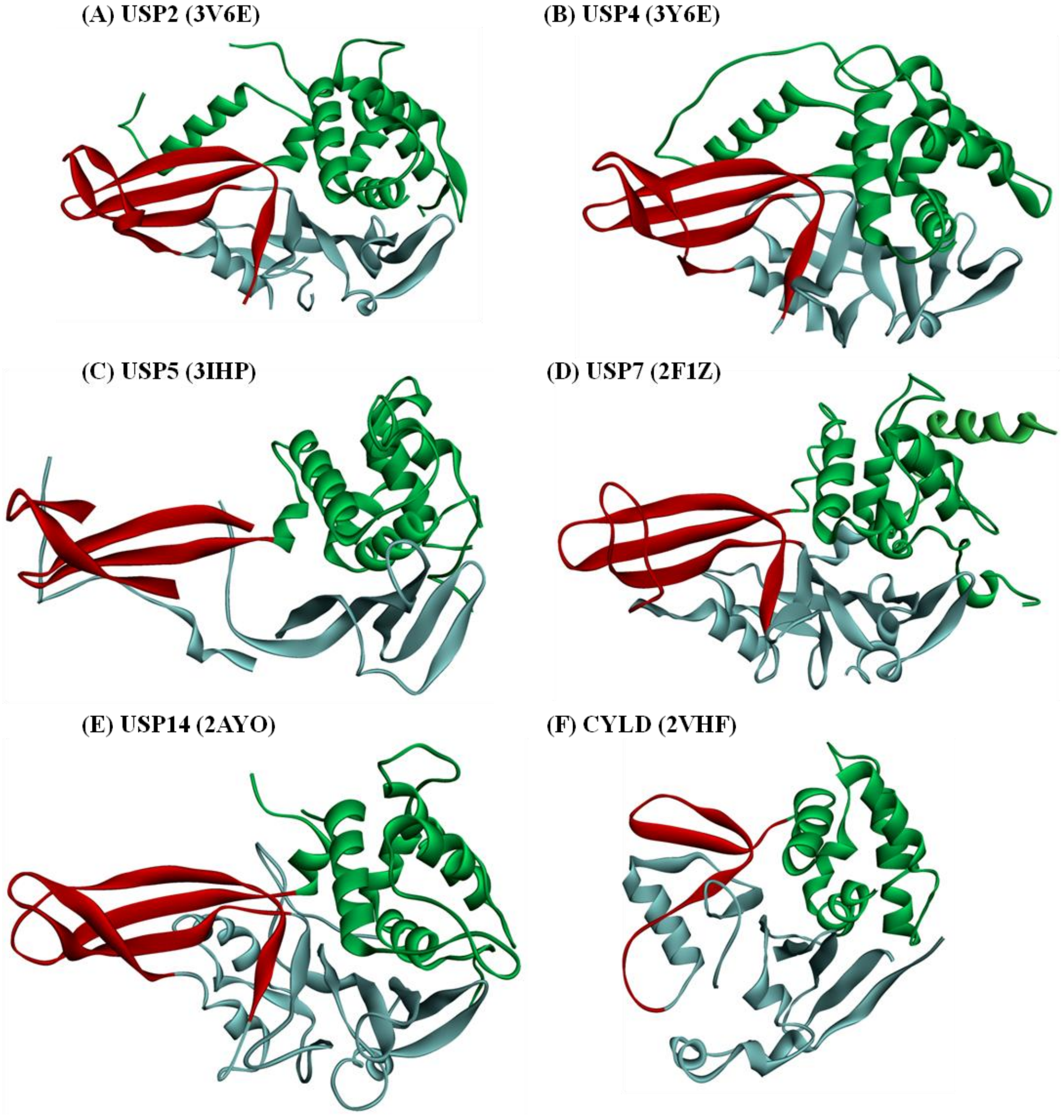


Figure 5.14. Structural comparison of C19 domains. Ribbon diagrams of C19 domain structure of the indicated human USPs (retrieved from the RCSB database) are shown. The characteristic subdomains of the C19 domain are shown: fingers (red), palm (cyan) and thumb (green). Note: the shortening of all the subdomains in CYLD compared to other structures. PDBids are indicated in brackets.

5.8. Gene expression profile of USPs in human

The observed expansion and subsequent retention of USPs across the animal lineages could be the function of extensive novelties in organs and tissue, that is concurrent with the evolution of taxonomic lineages examined in this study. Thus to explore the tissue specificity of different human encoded USPs, the BIOGPS database was used to note the expression pattern of those proteins in different anatomical regions of the body. BIOGPS is a public database that is built upon the information acquired from RNAs derived from 79 human tissues (Su *et al*., 2004). The relative quantity of gene specific RNA in the examined tissue is inferred as a measure of gene expression and the data is presented in graph, where bars represent the relative expression of the query gene, however this does not necessarily reflect the protein levels. For clarity the observations are tabulated in Table 5.6, where genes are categorized on the basis of relative level of expression. In total, relatively high levels of USP expression was noticed in tissues associated with the immune system, vasculature and nervous system. For example the data show that USP1 is expressed at 10 fold higher levels than the average USP1 level in CD34 cells. The CD34 (cluster of differentiation 34) molecule belongs to a family of single pass transmembrane sialomucin that is expressed on hematopoietic stem cells and vascular associated cells (Nielsen and McNagny, 2008) and certain cancer cell types (Casey *et al*., 2006; Ney *et al*., 2007; Nielsen and McNagny, 2008; Somasiri *et al*., 2004). USP1 plays a central role in the post translational regulation of proteins involved in Fanconi anaemia pathway for DNA repair (Murai *et al*., 2011; Nijman *et al*., 2005a) and translesion synthesis (Huang *et al*., 2004). USP3 shows more than 10 fold higher expression in CD33+ myeloid cells and CD14+ monocytes compared to mean expression in the examined tissues. To date the best characterized function of USP3 is its ability to deubiquitinate histones (H2A and H2B) which in turn facilitates the progression of the cell cycle (Nicassio *et al*., 2007). However, the functional relevance of high levels of USP3 in myeloid cells and monocytes is not known. USP7 RNA is found at highest level (more than 10 fold) in CD71+ early erythroid cells to the average level of expression among the 79 tissues examined. USP7 is a multifunctional protein and it is widely known for its role in the regulation of the tumour suppressor protein p53 and PML bodies (Hu *et al*., 2006), as well as transcription coupled nucleotide excision repair (Schwertman *et al*., 2013). CYLD, the only member of group 16, is expressed at more than 10 fold and 3 fold above average in CD4+Tcells and CD8+Tcells respectively. This is consistent with its role in the regulation of T cell development and function (Reissig *et al*., 2012).

GRPs	USPs	Gene Expression		
		≥10x	≥3x	>Median
1	**1**	CD34	CD71EE, CD105+EC, 721BLB, CD19B, BDCA4DC, CD56+NK	Testis
	12	0	CD71EE, Colon, Pineal Night (PN), Pineal Day (PD), Pre-Frontal Cortex (PFC)	CD105+EC
	46	0	Amygdala (AMY)	PN,PD, Retina
	35	ND	ND	ND
	38	ND	ND	ND
2	**2**	0	0	0
	21	0	CD34+	CD4T, CD8T
	8	0	PN, PD, CD19B, BDCA4DC	CD4T, CD8T, Thyroid
	50	0	0	Skin
	36	0	CD19B, BDCA4DC	CD4T, CD8T, Heart
	42	0	0	0
	17	0	0	Heart, Liver, Pancreas
3	**3**	CD33My CD14Mo	CD4T, CD8T, CD56NK, CD19B, BDCA4DC, 721BLB	0
	33	0	PN, PD, AMY, PFC	CD4T, CD8T, CD19B, BDCA4DC, Hypothalamus, Thyroid
	20	CD8T	CD71EE, CD4T, CD56+NK	0
	16	0	CD33My, CD19B, BDCA4DC, CD56+NK, 721BLB	0
	45	0	0	0
	44	0	721BLB	Testis
	49	0	0	Heart, Trigeminal ganglion (TGG), Parietal lobe, Super Cervical Ganglion (SCG)
	22	0	0	Heart, TGG
	51	0	0	0
	27	0	SCG	0
4	**4**	0	CD56+NK	CD34+
	15	CD71EE	CD105+EC, CD14Mo	0
	11	0	PN, PD, PFC, Temporal Lobe, Hypothalamus	Retina, Olfactory Lobe
	32	0	CD33My, Testis	CD71EE
	19	0	TGG	0
	43	0	--	0
	31	0	0	TGG, Testis
5	**5**	0	0	721BLB
	13	Skeletal Muscles	PN, 721BLB	0
6	**6**	0	Testis	0
7	**7**	CD71EE	CD105+EE, 721BLB, CD19B, CD56+NK, CD33My, CD14M	PN, PD, PFC
	47	0	CD56+NK, CD4T, CD8T, PFC	0
	40	0	0	Prostate
	34	0	0	PFC, SCG, Pons, Prostate
	24	0	0	CD4T, SCG
	9X	0	721BLB	PN, AMY
	9Y	0	PN, PD	0
	48	0	TGG, SCG	0
	18	0	721BLB	0
	41	0	0	0
8	**10**	0	0	721BLB, BDCA4DC
9	**14**	0	CD34+, 721BLB, AMY, PFC, Prostate	0
10	**25**	0	721BLB, BDCA4DC, CD4T, CD8T, CD56+NK, CD33My, Testis	0
	28	CD56NK	0	0
11	**26**	0	0	0
	29	0	0	0
	37	0	0	0
12	**30**	0	0	0
13	**39**	0	0	CD34+, CD105+EC, 721BLB
14	**52**	0	721BLB, CD4T, CD8T	CD34+
15	**53**	0	0	Heart
	54	0	0	0
16	**CYLD**	CD4T	CD8T, PD, PN	PFC

Table 5.6. Gene expression profile of human USPs. The Gene expression profile of human USPs was retrieved from BIOGPS. Cells and tissues are coloured according to the systems and/or associated functions. Expression level is shown in comparison to the average (of tissues analysed) and categorised as >10x, >3x or >Median. Key: Immune system (green), vasculature (red) and nervous system (blue), glands (pink), reproductive (brown), others (light brown).

Several gene pairs that segregated at the root of vertebrate divergence from a single lineal ancestral gene show difference in their tissue expression pattern in humans. For example, USP12 shows higher than the tissue average expression in immature erythroid cells (CD71+EE) and in pineal gland tissue, while USP46 (the closest paralogue of USP12) shows highest RNA abundance in the amygdala, part of the limbic system in the brain modulating behavioural responses (Amunts *et al*., 2005). Interestingly, it has been recently observed that USP46 knockout mice behave differently compared to wild type mice in response to antidepressant drug, Nitrazepam (Imai *et al*., 2012), indicative of concurrent presence of tissue specificity and functional divergence between USP12 and USP46. Similarly, USP2 is expressed almost uniformly in all the examined tissues while >3 fold higher expression of its closest paralogue USP21 was observed in hematopoietic stem cells (CD34+). USP33 and USP20 are phylogenetically the closest paralogues of group 3, the former is expressed predominantly in tissues of the nervous system and the latter in cells involved in the cell mediated immunity (T cells). Despite the difference in tissue specific expression patterns USP20 and USP33 show functional similarities as both are associated with recycling of β adrenergic receptor (Berthouze *et al*., 2009). Similarly, the expression patterns of USP4, USP11 and USP15 of group 4 differ, where USP4 and USP15 are predominantly expressed in the vasculature and/or immune related cells, whereas higher expression of USP11 was observed in different regions of the central nervous system. Interestingly, USP4 and USP11 negatively regulate TNFα induced NFκB activation (Fan *et al*., 2011; Sun *et al*., 2010), by contrast USP15 promotes TGFβ cell signalling, which can support oncogenesis (Eichhorn *et al*., 2012), suggesting functional differences between these closely related paralogues. USP5 and USP13 are the two paralogues of group 5, where USP5 is almost uniformly expressed in the compared tissues whereas USP13 shows >10 fold average expression in skeletal muscle. Though a role of USP13 has been demonstrated in cell proliferation and regulation of gene expression (Zhao *et al*., 2011), the direct physiological relevance of significantly high expression of USP13 in skeletal muscle is unknown.

The data indicate that many USPs are expressed in cells with an immune related function and several others in cells of the nervous system and vasculature. Interestingly, the USPs which emerged at the origin of vertebrates also demonstrate divergence in their tissue specific expression and molecular functions, showing a lack of redundancy and hence providing an explanation for their retention.

5.9. Protein interaction network analyses of USPs

In order to further our investigation regarding the functional divergence of USPs, each human USP was explored using STRINGv9.1 data (Table 5.7). STRING is a public database that provides information regarding known and predicted, direct (physical) or indirect (functional) protein-protein interactions. STRING retrieves the protein-protein association information from four different sources: genomic context, high throughput screening, co-expression and previous knowledge (text mining) (Franceschini *et al*., 2013). The server offers several levels of statistical robustness from relaxed to highest confidence. In this study, associations with only high confidence score (≥0.7) are considered.

As reflected by association with ubiquitin C (UBC), regulation of protein turn over appeared as a common feature for almost all USPs. Other common associations detected by the protein network analysis are the association of USPs with proteins involved in DNA repair. For example, USP1 association was observed with FANC1, FANCD2 and ATAD5, which are important components of the DNA damage response and replication (Lee *et al*., 2013; Murai *et al*., 2011). Similarly, association between KIAA1530 (UVS SA) and BRCA with USP7 was found, both these USP partner proteins are involved in nucleotide excision repair response (Deng *et al*., 2003a; Schewertman *et al*., 2013). Another common feature to which most USPs seems to be associated with is regulation of gene expression, for example, association of many USPs was identified with multiple components of the STAGA complex, which is a chromatin acetylating transcription co-activator (Martinez *et al*., 2001). Similarly association of USP7 was observed with several ubiquitin ligases (including RING1) and the chromatin binding factor (ATXN1), reflecting its role in the regulation of gene expression. Many USPs were also found to be associated with molecules that are connected with apoptosis. Among these, several apoptosis associated factors such as p53 and FOXO4 associate with USP7. Similarly, another paralogue of group 7, USP9X, establishes a network with proteins involved in apoptosis. This is consistent with empirical observations as these and many USPs are known to contribute in the regulation of several key factors of apoptosis (Ramakrishna *et al*., 2011).

Several phylogenetically close paralogues show differences in the known and predicted binding/associated partners, which in turn reflect their functional divergence. For example USP20 and USP33, the closest paralogues of group 3, differ in their network analysis. USP20 establishes a network with molecule involved in gene regulation and angiogenesis (ERG) (Birdsey *et al*., 2012) while USP33 associates with proteins involved in the development of the nervous system such as ROBO1 (Long *et al*., 2004) and DIO2 (Guo *et al*., 2004). USP25 and USP28 are the only two paralogues of group 10, and they differ in

GRPs	USPs	Partner Proteins
1	1	UBC, WDR48, FANCD2, FANCI, ATAD5, ZBTB352
	12	WDR20, WDR48, DMWD
	46	WDR20, WDR48, UBC, PHLPP1, DMWD
	35	NDA
	38	NDA
2	2	UBC, UBA52, FASN
	21	UBB, UBC, UBA52, RIPK1, KCTD13, KCTD10, BTBD9
	8	UBC, UBB, UBA52, RNF41, RNF128, DNAJB6, BIRC6, OTUB1, KIF23, EGFR, AKT1, EPS15, GRB2, STAMs, CHMP1A, CHMP4C, C10orf2, *C18orf2*
	50	IMP5
	36	UBC, CDK4, DNAH14, DNAH5, DNAH12, DNAH19, DYNC1H1
	42	NDA
	17	SUDS3
3	3	UBC, EIF3CL, H2A, H2B
	33	VHL, ATG3, ROBO1, DIO2, OS9, ARRB2
	20	VHL, USP16, ERG
	16	UBC, USP20, FUCA1, MARK1, MARK2, MARK3, MARK4, PRKC1
	45	NDA
	44	NDA
	49	PPT1
	22	UBC, KAT2A, KAT2B, TRRAP, ATXNL3, ENY2, STAGA-Trans-HAT Complex, SAP130, Sin3A, FAM48A, H2A, H2B
	51	NDA
	27	NDA
4	4	UBC, SART3, GRP, RB1, PRPF3, LSM2
	15	UBB, UBC, UBA52, PSMD7, UCHL5, GRP, SART3, LSM2, SMAD7, TGFβR1, TGFβR2, TGFβ1
	11	UBC, USP7, BRCA2, WRNIP1, TCEAL1, CBX8
	32	NDA
	19	UBC, BIRC2
	43	NDA
	31	NDA
5	5	UBC, UBA52
	13	UBC, UFDIL, SERPINCI, G6PD
6	6	NDA
7	7	UBC,UBA52, USP11, UHRF11, PSMA complex, PSMB complex, KIAA1530, BRCA1, p53, FOXO4, DAP6, MDM2, MDM4, DAXX, TRAF6, GMPS, DNMT1, CLSPN, PPMIG, RING1, RNF2, C14orf2, RNF220, SRF, BMI1, ATXN1, TSPYL4
	47	FBXL3, FBXL15, FBXO2, FBXL7, SCF-complex, SARS2
	40	NDA
	34	NDA
	24	UBC
	9X	UBC, ITCH, BIRC5, MCL1, HUWE1, TMEM49, MTOR, NUAK1, MLLT4, CTNNB1, MARK4
	9Y	NDA
	48	Lys6-D, PRUNE
	18	ISG15, IFNAR2
	41	NDA
8	10	UBC,G3BP1, SNX3, CFTR
9	14	UBC, UBA52, PSMA complex, PSMB complex, PSMD complex, ERG, CASP1, CDKL2, KIR2DL3, CD68
10	25	UBC, SUMO2, SUMO3, KLH13, MYBC1
	28	UBC, SUMO2, FBXW7,TP53BP1, CLSPN, MDC1, MYC
11	26	NDA
	29	NDA
	37	NDA
12	30	NDA
13	39	UBC, SART3, SLC25A4
14	52	PAN3
15	53	NDA
	54	CHMP2A, CHMP4A, CHMP6
16	CYLD	UBC, HDAC6, PLK1, TRAF2, TRAF6, RIPK, SQSTM1, TRAIP, OPTN, IKBkG, IKBkE, BCL3, TBK1, DVL1, DDX58, LCK

Key: Protein turnover, DNA repair, Transcription, Cell cycle and division, Apoptosis, G protein signalling, Embryogenesis, General metabolism, TNF signalling, Channel protein, ER trafficking and protein degradation, EGFR signalling, TGFβ signalling, *Unknown function*, Cilliary motility, Protein synthesis, Cytoskeleton, RNA processing, NFkB signalling, Cell adhesion, WNT signalling, Immunity, Musculature, NDA: no data available

Table 5.7. Known and/or predicted protein binding partners of human USPs. Partner protein interactions of USPs were examined using STRING v9.0. USP partner proteins predicted with a high confidence level (≥0.7) are included. Interacting molecules are coloured differently according to the main associated functions (see key). USP paralogue boxes are shaded: (light brown) vertebrate specific, (light blue) mammalian specific and primate specific (light green). Abbreviations of all molecules are provided in Appendix V.

their association in the protein network analysis. USP25 is found associated with KLH13, a protein involved in cytokinesis (Sumara *et al*., 2007) while USP28 establishes a network with CLSPN which is associated with the DNA damage response in the cell cycle (Bassermann *et al*., 2008).

In total, the differences in the protein network of USPs demonstrate the subfunctionalization between distantly and closely related USP paralogues. Despite retaining the core function of regulating protein turnover, these genes contribute to a variety of functions, reflected by the diversity of their molecular partners.

5.10. Summary of findings

- In total 55 paralogues of USPs are encoded by the human genome (excluding multiple paralogues of USP17), of these, most homologues are present in all the vertebrates examined. Some exceptions in this regard are USP6 and USP41 (present only in greater apes), USP17, USP51, USP26 and USP29 (present only in eutherian mammals).
- The Ensembl paralogy pipeline divides all human USP homologues into 16 paralogous groups and in most cases homologues of at least one paralogue from each group were found in slime mould and/or chlamydomonas suggesting a deep ancestral root of these USP homologues.
- Only two homologues of C19 domain containing proteins were identified in bacteria (*Candidatus amoebophilus asiaticus*), out of the >4500 prokaryotic genomes examined. This might suggest these have been acquired by horizontal gene transfer from eukaryotes to prokaryotes.
- C19 domain based phylogenetic reconstruction and data mining demonstrated a rapid expansion of the USPs genes at the base of the vertebrate tree.
- Genomic synteny of certain closely related paralogues points to the role of the whole genome duplication in the expansion of many vertebrate specific USPs.
- The origin of many eutherian specific USPs could be explained as a result of intra chromosomal segmental duplication.
- USPs paralogues of most organisms show a noticeable variation in the domain architecture and protein length.
- The domain architecture tends to be conserved between the orthologues of most USPs, supporting the idea that the C19 domain based phylogeny represents the evolutionary relationship of the full length gene.
- An increment in the domain combinations in the USP proteins was observed with the increase in organism complexity.

- Despite the sequence variation, strong structural conservation between the C19 peptidase domains of different USPs reflects common ancestry and functionality.
- The retention of the many USPs paralogues and/or ohnologues could be explained in terms of the divergence in the tissue specificity and functionality as observed by the comparison of expression data and protein association network.
- The origin of certain USPs is parallel to the origin of the substrate or partner protein and associated molecular pathways.

5.11. Discussion

The phylogenomic analyses of the USPs conducted here delineates time points in relation to the origin of different USP homologues in animal lineages and point to the underlying mechanisms of their expansion and subsequent retention.

Origin of USPs

Datamining and phylogenetic analyses show that at least one homologue of nearly all the USP paralogous groups existed in slime mould (protozoa), which indicate that the last common ancestor of animalia (including protozoa and metazoa) had at least one antecedental gene of most of the paralogous groups of USPs. From these expansion and diversification at different speciation points lead to the formation of the extant array of USPs in humans. Interestingly, 11 peptidase C19 domain containing proteins are present in chlamydomonas and showing orthologous relationship with paralogues of group 1, 2, 4, 5, 7, 8, 9, 13 and 14. This suggests the origin of respective USP groups before the animal-plant split. The number of USP homologues increases from chlamydomonas (green algae) to *Arabidopsis thaliana* (eudicot), suggesting the possible convergent expansion of USPs in plants and animal kingdom. It is of great interest for evolutionary studies to explore the potential array of USPs present in the common ancestor of all eukaryotes by comparing plant and animals USPs. However, given the 2,738 million years of independent evolution, difference in the nucleotide substitution rate (Buckley and Cunningham, 2002) and variable gene death (Roy *et al*., 2009), such an endeavour demands an extensive phylogenetic analysis beyond this study. Out of over 4,500 bacterial and archaeal genomes screened, the presence of the C19 domain in only one bacterial species (*Candidatus amoebophilus asiaticus*) is surprising. Intriguingly, *C. Amoebophilus asiaticus* is an obligate intracellular symbiont of amoeba (Schmitz-Esser *et al*., 2008), this raises the possibility that C19 domain may have been acquired as a result of horizontal gene transfer (HGT) from the eukaryotic cell (amoeba) and not originated in the prokaryotic genome. Alternatively, the peptidase C19 domain had indeed originated in the prokaryotes and nearly all bacterial species had lost it after the prokaryote eukaryote split. However, the genome sequence analysis of *C. Amoebophilus asiaticus* shows the presence of over 50 foreign genes in the bacterial genome which also include two genes that share 27% sequence identity with ubiquitin carboxyl terminal hydrolase of *Trichomonas vaginalis*, a protozoan (Schimitz-Esser *et al*., 2010). This strongly supports the bacteria may have acquired the C19 domain as result of horizontal gene transfer. This brings a caveat in the phylogenetic analysis as all trees were rooted using *C. Amoebophilus asiaticus* C19 domain sequence. However, in all trees the sequence automatically rooted out from the homologues found in eukaryotes. Additionally, the unrooted trees also retained the topology observed in the rooted trees.

Structural similarities indicate the common ancestry

Owing to the limited number of structural folds and strict evolutionary constraints on the folding pattern, protein structures are often considered as better evolutionary markers than the gene or protein sequences, especially to explore distant ancestral relationships (Agarwal *et al*., 2009; Liu *et al*., 2004; Scheeff and Bourne, 2005). Thereby, strong structural conservation in a protein reflects common ancestry despite the lack of support from traditional sequence based approaches (Agarwal *et al*., 2009; Scheeff and Bourne, 2005). Comparison of the six C19 domains (examined here) revealed strong structural similarities between the peptidase C19 domains of USPs regardless of their association with different paralogous groups. This reflects that different variants of the C19 domain may have evolved from a common ancestor, however, over time, the sequences diverged considerably but with limited effect on the overall domain architecture. Alternatively, the possibility that these structures developed through convergent evolution cannot be completely discounted (Bukhari and Caetano-Anolles, 2013; Tomii *et al*., 2012).

Origin of USP paralogous groups

The phylogenetic analysis reveals that nearly all USP paralogous groups emerged before the origin of metazoa. Exceptions in this regard are the USP paralogous groups 6, 10, 11, 12, 15 and 16, which incorporate relatively few USP homologues (1-3). Among those, the origin of groups 10, 12, 15 and 16 is placed with the origin of coelomates (nearly 1000 MYA). One homologue of group 11 (comprising USP26,29,37) was identified in hydra and the others coelomates suggesting its origin lies in the common ancestor of metazoans. USP6 (group 6) is composed of two protein domains, a TBC domain and C19. Only homologues found in humans, chimpanzee and gorilla possess both domains, while homologues found in all other animals have only the TBC domain. This suggests a recent acquisition of the C19 domain in the common ancestor of great apes probably from USP32 (based on the phylogenetic analysis). Similarly, CYLD (the only human USP homologue of group 16) is comprised of two protein domains: CAP-GLY and C19N. The CAP-GLY domain was identified in genes of both slime mould and hydra but without a C19 domain. This points to the origin of CYLD by domain acquisition with the origin of coelomates.

Owing to the very low sequence identity between the members of different paralogous groups, sequence based phylogeny (by composite tree reconstruction of all USPs) lacks resolution (data not shown). However, the evolutionary relationship between different USP paralogous groups could be investigated using structural based phylogenetic methodologies. In addition some clues have been gathered from individual species trees, which points to a common ancestry for group 2, 3 and 4. Similarly group 1 and 7 show some linkage patterns as do groups 5 and 10.

Birth and death of USPs in metazoa and role of whole genome duplications*; *Model of USPs Evolution

Given the extent of sequence divergence between USP paralogues, the potential gene loss in the species analysed and the species bias represented in the sequenced genomes, it is difficult to develop a reliable composite phylogeny of all paralogous USP groups together. Therefore phylogenetic trees were reconstructed separately for each paralogous group using the C19 domain and based on their topology, and the distribution of vertebrate and non vertebrate homologues, a model tree has schematically drawn (Figure 5.15). Although the data indicate that most USP paralogous groups originated before the protozoan-metazoan split (1184 MYA) and/or possibly before animal-plant split (1369 MYA), genes within several paralogous groups underwent subsequent expansion at different time points particularly during the emergence of coelomates and vertebrates. The latter time is also marked for the proposed two whole genome duplication events occurred between the divergence the vertebrates and chordate-vertebrate split (722 MYA) (Ohno *et al*., 1970; Holland *et al*., 1994). It has been proposed that the first whole genome duplication event (1R) occurred between the divergence chordates and divergence of jawless fishes while the second whole genome duplication event (2R) happened between jawless fishes and bony fishes (Escriva *et al*., 2002; Kuraku *et al*., 2009). Moreover, after the fish-tetrapod split, the common ancestor of ray finned fishes (400 MYA) also underwent a whole genome duplication event (3R) (Amores *et al*., 1998; Postlethwait *et al*., 1998). If the genes produced via these genome duplication events (referred to as ohnologues) survived during the course of evolution, each of such duplication events would have resulted in the 2 fold increase in the number of antecedental genes, thereby 4 and 8 copies of USPs would be present in tetrapods and fishes respectively, against each lineal ancestral gene. However, consistent to the "Birth and Death" model of gene evolution (Nei and Rooney, 2005),

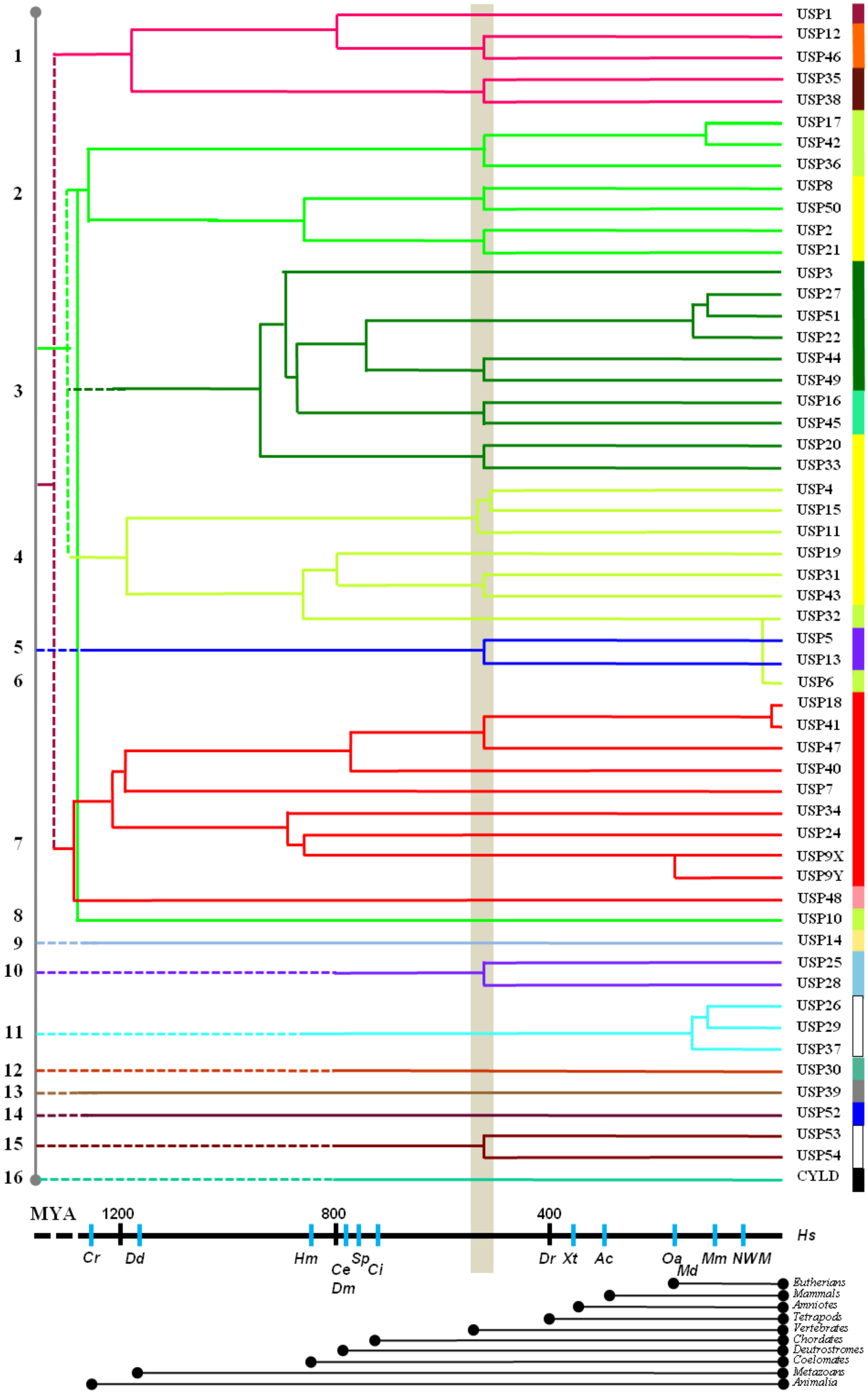

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
USP1
USP12
USP46
USP35
USP38
USP17
USP42
USP36
USP8
USP50
USP2
USP21
USP3
USP27
USP51
USP22
USP44
USP49
USP16
USP45
USP20
USP33
USP4
USP15
USP11
USP19
USP31
USP43
USP32
USP5
USP13
USP6
USP18
USP41
USP47
USP40
USP7
USP34
USP24
USP9X
USP9Y
USP48
USP10
USP14
USP25
USP28
USP26
USP29
USP37
USP30
USP39
USP52
USP53
USP54
CYLD
MYA
1200
800
400
Hs
Cr
Dd
Hm
Ce
Sp
Ci
Dm
Dr
Xt
Ac
Oa
Md
Mm
NWM
Eutherians
Mammals
Amniotes
Tetrapods
Vertebrates
Chordates
Deutrostromes
Coelomates
Metazoans
Animalia

Figure 5.15. Model for evolution of USPs in animalia. Evolutionary relationship of USPs were explored using the maximum and likelihood method. The tree is manually redrawn to collate the informations gathered by the phylogenetic reconstructions of individual paralogous groups regarding distribution of the vertebrate and nonvertebrate homologues in the tree. Branches corresponding to each gene is coloured differently to represent their associated paralogous group and drawn to scale with estimated evolutionary timeline (shown below). Major taxonomic association of the animals are shown below. The light bar represents to the evolutionary period where two events of whole genome duplications have been proposed. Organisms are abbreviated as defined in section 5.3.

neutral drift in evolution quickly eliminated most of the duplicated paralogues due to redundancy. Alternatively, the daughter genes could undergo the process of fixation which leads to the sub-functionalization and/or neo-functionalization (Innan and Kondrashov, 2010). In agreement, the present phylogenomic analyses are indicative of a “Birth and Death” patterns of evolution in the USP genes. More than half of the potential USP ohnologues which emerged after 1R/2R have succumbed to gene death in vertebrates, whereas the remainder underwent a process of fixation and preservation in the population. Genome analysis of lancelet (*Branchiostoma floridae*) has also shown that only one quarter of the duplicated paralogues retained in the human gene families and much smaller fraction of these are ohnologues (Putnam *et al*., 2008). Although the proposed 3R event occurred in the common ancestor of ray finned fishes, no evidence of additional USPs was found among the fish genomes, suggesting extensive gene death in USP ohnologues generated by the fish specific WGD. A similar extent of gene death was observed through the phylogenetic analysis of the GATA gene family, where despite the retention of ohnologues originating from 1R/2R, all teleosts fishes have lost all the ohnologues which had emerged as a result of 3R (Gillis *et al*., 2009). Conversely, in GH18 family genes, evidence for the relics of 3R have been presented (discussed in chapter 6) (Hussain and Wilson, 2013). Nevertheless, phylogenetic reconstruction shows 13 pairs of vertebrate specific USP homologues (including the triplet of USP4, USP11 and USP15) share a lineal ancestry with a single non vertebrate gene. Out of these, the genomic loci of 7 pairs of phylogenetically linked genes exhibit other paralogous gene pairs in their proximity. Furthermore, based on paralogous chromosomes regions determined in the human genome, relics of both WGDs have been partially mapped (Dehal and Boore, 2005; Nakatani *et al*., 2007), only four of the vertebrate specific USPs map to these regions which include: USP36/42, USP20/33, USP4/11/15 and USP5/13. Both these observations (presence of paralogons and position of the genes on human genome) support the model of USP gene expansion in early vertebrates via whole genome duplication events. It is noteworthy that paralogons of USPs are discontinuous in nature as in many cases non paralogous genes disrupt the stretch of paralogous genes. Additionally, 5 vertebrate specific and phylogenetically linked USP gene pairs shows no syntenic support in connection to WGDs in the compared region and with the loci mapped for WGD while one pair USP8/50 likely to be originated by gene duplication. These observations may be due to the limit placed in this comparison of 10 genes on each side of the locus. However, the same threshold 10 genes was found suitable in the previous studies using a proximate gene pair method for identifying true syntenic regions which had arisen by WGD (Panopulo *et al*., 2003; Hufton *et al*., 2008). Alternatively and perhaps more likely, the lack of synteny among the vertebrate specific USP pairs is the result of extensive post WGD gene death and/or genomic rearrangement, thereby losing the genomic synteny, as proposed by Nakatani *et al* (2007) and Hufton *et al* (2008). Moreover, it is suggested that the WGD events themselves increased the rate of genomic rearrangement (Otto, 2007), however, later studies have shown no cause and effect correlation between WGDs and genomic rearrangements but instead demonstrated an increased rate of synteny loss in early tetrapod lineages (Huffton *et al*. 2008). Counterintuitively, existence of USP paralogons could also be inferred as a result of large scale segmental duplication in the ancestral genome followed by genetic rearrangement over time. Finally, as it is assumed that the genome duplication events occurred between the emergence of chordates and early vertebrates, inclusion of homologues from organisms that reflect this time frame, such as cartilaginous fishes (once the genomic assembly is available) and lamprey, may further refine the timeline for the expansion of USPs at the root of vertebrates.

Examining these limited number of species revealed no evidence of expansion in USPs with the origin of mammals however, some paralogues originated after the marsupial-eutherian split (162 MYA). In this regard one interesting case is of USP17, where multiple homologues are arranged on human chromosome 4p and 8q. A previous phylogenetic study by Burrows *et al*., (2010) demonstrated the clustering of USP17 paralogues in a single clade, suggesting that these duplications happened after the speciation events in the eutherian mammals. However, given the relatively recent gene expansion of USP17, it is possible that the gene sequences, especially of the C19 domain, have not diverged enough to distinguish between paralogues and orthologues. It has been suggested that most breakpoints in the human and mouse genomes occur close to tandem gene duplications or large segmental duplications (Armengol *et al*., 2005). Hence it is tempting to speculate that the tandem duplication of USP17 lead to chromosomal breakage and relocation, resulting in the presence of USP17 homologues on the two chromosomes (4p and 8q). 9 out of 32 copies of USP17 homologues are either catalytically inactive or exist as pseudogenes. Similarly, two human USPs, USP50 and USP54, are also catalytically inactive. However, USP50 is catalytically active in all vertebrates examined except human suggesting that loss of the enzymatic activity in USP50 is relatively a recent evolutionary event.

With phylogenetic support present, homologues adjacently located on a chromosome can be the product of gene duplication (Hussain and Wilson, 2013; Cridland *et al*., 2012). Among the USPs, except for USP17 homologues (eutherian specific) USP8 & USP50 (vertebrate specific), which are located adjacent on a single chromosome, no other USP pairs are proximally present indicating a limited role of gene duplication in the expansion of USPs in vertebrates. Some of the USP genes such as USP18&USP41 and USP27&51 are phylogenetically linked and found on different but paralogous regions of the same chromosome, suggesting these pairs of genes originated via segmental duplication. It is interesting to note that most of those genes have limited distribution in eutherian mammals (USP27, USP51) or two species of primates: human and chimpanzee (USP41). Of note, earlier observations have shown an increased proportion of interspersed segmental duplications within the genomes of humans and great apes (Marques-Bonet *et al*., 2009; Zhou and Mishra, 2005) compared to other mammals. Though segmental duplication is generally associated with the genomic instability (Emanuel and Shaikh, 2001) recently it has been shown to underlie the emergence of promoters for LRRC37 gene family (Bekpen *et al*., 2012), suggesting segmental duplication has also contributed in the expansion of other gene families.

In summary, the USP gene family has undergone extensive expansion at the root of the vertebrates and limited expansion with the origin of eutherian mammals and great apes. Several genetic forces such as genome duplication, segmental duplication, gene duplication and domain acquisition have contributed at different time points in the expansion of the USPs in the extant vertebrates and eutherian mammals.

Conflict between gene tree and species tree

Given the considerable sequential divergence among the USPs, it is not surprising that positioning of certain branches and nodes in the phylogenetic reconstructions are not super imposable upon the established species tree. Discordances between multigene trees and the species tree are not uncommon in phylogenetic analyses (Degnan and Rosenberg, 2006; 2009; Nichols, 2001). Discordances between the species and gene trees, also termed anomalous gene trees; (AGT; Degnan and Rosenberg, 2006; 2009) are more evident in paralogues with limited distribution. For instance USP17 homologues are only present in eutherian mammals and yet it outgroups vertebrate USP36 and USP42 in a single monophyletic clade (Figure 5.3). Similarly, the arrangement of the orthologues with in clade/subclades also deviates from the established speciation events. Several explanations could be proposed to account for these inconsistencies: 1) extensive loss of the gene causes the misplacement of the clade or subclades that deviate from the established speciation events; 2) bias in the available and/or analysed genome sequences of animals (more mammalian genomes have been sequenced/analysed than other taxonomic lineages); 3) loss of phylogenetic signals (less informative substitutions) because of the usage of limited sequence length (for example C19 domain sequences instead of full gene sequences); 4) heterogeneity in the substitution rate between the compared homologues. 5) AGT can also result from the fast adaptive radiation resulting in divergence in rapid succession.

Peptidase C19 is the only domain that is consistently present in other wise highly divergent USPs. Moreover, for multidomains proteins, conserved domain sequences are often employed successfully to examine the evolutionary history of the corresponding genes (Alvarez-Venegas and Avramova, 2012; Rojas *et al*., 2012). This and the observation obtained from the phylogenetic analysis (most clades with >90% bootstrap value) suggests that C19 domain sequence provides sufficient informative signal to reconstruct the phylogenetic relationship of USPs. Finally variation in the nucleotide substitution rate, as reflected by the branch length of individual clades within the tree and rapid adaptive radiation could be accounted to reason the AGT in the present phylogenetic reconstructions.

Domain diversity and organismal complexity

Protein domains are considered as a distinct evolutionary unit and it has been suggested that multidomain architectures reflect organismal complexity (Konnin *et al*., 2002; Vogel *et al*., 2004). However, other lines of evidence suggest that domain promiscuity is independent of complexity of the organisms and similar multidomain architecture could evolved convergently, independent of the evolutionary position of the organism (Forslund *et al*., 2008). The domain analysis of USP homologues in species associated with key phylogenetic points demonstrates a trend of incremental complexity. While retaining the inherited domain conformations, new variants of peptidase C19 domain and domain combinations arose in USPs during the transition from a simpler form to relatively more complex form. Most domains combinations (such as MATH-C19C-ICP0 and WD40-C19P-PAN2 etc) are stably inherited from protozoa to primates. This strong conservation in domain combinations between orthologues suggests a functional conservation of USPs orthologues across animal lineages. By contrast, the domain versatility between paralogues reflects functional divergence. Biological mechanisms leading to the generation of new domains or domain combinations are not fully understood. However, processes such as gene fusion and loss/gain of protein domains or shuffling of protein domains are often proposed as major contributing factors in this regard (Bork, 1991; Chothia and Gerstein, 1997). Moreover, progressive folds do evolve over time in domains, resulting in the variations within the protein domains and even origin of new domains (Grishin, 2001; Scheef and Bourne, 2005).

Although domain combinations among most USP orthologues remains fixed throughout the evolutionary history, there are a few exceptions in this regard. For example the rhodanese domain and domain of unknown function (DUF) are present in all orthologues of USP8 except of hydra (XP002164031) which points to their acquisition in the USPs in early coelomates. However, the rhodanese domain is frequently found in other proteins of bacteria, fungi and plants (reviewed in Bordo and Bork, 2002), demonstrating the prokaryotic origin of the domain and likely subsequent assembly (via domain acquisition) into USP8 in the common ancestor of coelomates. Alternatively, the domain could have been present in the ancestral USP8 and subsequently lost in hydra after its divergence from the common ancestor. Typically, the rhodanese domain is catalytically active and composed of two identically folded sub-domains with very weak (13%-21%) sequence identity (Bordo *et al*., 2000). In USP8 orthologues, only the N-terminal half of the domain is present which lacks the catalytic activity and has been shown to interact with the E3 ligase NRDP1 (Avvakumov *et al*., 2006). Given that this sub domain is also absent in the USP8 orthologue in hydra, it is reasonable to suggest that USP8 orthologue may be functionally different from USP8 of other species. Similarly, Zn finger domain (Zf) are not present in slime mould and NV orthologues of USP16,45 and USP20,33 suggesting the origin of the DNA binding capacity of these proteins emerged with coelomates and vertebrates respectively. In certain orthologues C19 associated domains appeared to have duplicated over time during the evolution of animals. For instance the CYLD orthologue in fruit fly (NM164910) carries a single CAP-GLY with a C19N domain, whereas in all other deuterostome orthologues of CYLD there are 3 CAP-GLYdomains along with C19N, suggesting domain duplication occurred. The CAP-GLY domain is an 80 residue long protein module which is involved in microtubule organization and transportation of vesicles and organelles (Steinmetz and Akhmanova, 2008). Since the CYLD orthologue of fruit fly and USP19 orthologues of NV deuterostomes are yet to be functionally characterized, the significance of this domain duplication is not known. However, it is suggested to be a common feature in protein evolution (Nacher *et al*., 2010).

As Zf, UBA, DUSP and peptidase C19R domains (not mutually exclusive) are present in USPs of different paralogous groups, this indicates that domain combinations could be emerged several times. Alternatively it may indicate distant shared ancestry between those paralogous groups.

In summary, the phylogenomic and domain distribution analysis show the stable transition of multidomain architectures from the ancestral lineages (represented by slime mould) to the primates (humans), which reflect functional conservation between the orthologues. Moreover, domain comparison of USPs homologues shows an increase in the domain combination paralleling organismal complexity. Finally, it is important to note that USP proteins in humans show considerable variation in length and most of the protein regions are not annotated for any structural or biological significance.

USPs evolution and functional divergence

The retention of duplicated genes in an organism is indicative of functional divergence which in turn is reflected by domain diversity, expression profile and partner protein interaction of the corresponding genes (Innan and Kondrashov, 2010; Lynch and Conery, 2000).

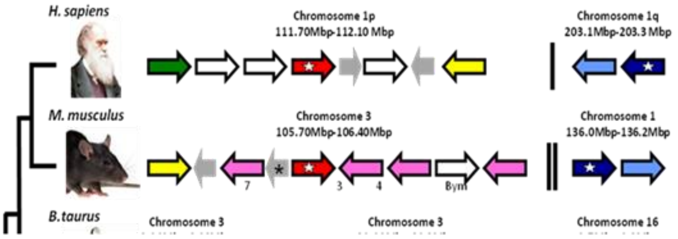
Homologues found in slime mould include: USP12, USP46, USP36, USP7, USP16, USP15, USP47, USP40, USP5, USP10, USP14, USP39, USP48 and USP52. Most USPs present in protozoa are associated with core eukaryotic functions such as DNA repair (USP7, USP47 etc) (Parson *et al*., 2012; Sarasin, 2012), RNA processing (USP39, USP52) (Bett *et al*., 2013; Rios *et al*., 2011) and cell division (USP39 and USP16) (van Leuken *et al*., 2008; Joo *et al*., 2007). All of these functions have shown considerable transition as life evolved from prokaryotes to eukaryotes. Parallel origin of these USPs with eukaryotic cells implicates the role of USPs in the evolution of complex molecular machinery in eukaryotes.

Lineal ancestry of many human USPs were traced back prior to the origin of vertebrates. Molecular partners of these USP homologues suggest that most of the homologues that had originated with the origin of metazoans are associated with DNA repair, cell division, induction of apoptosis and cell cycle control. For example USP22 and USP3 deubiquitinate H2A and H2B and contribute in the cell cycle progression (Zhang *et al*., 2008), USP37 once phosphorylated with CDK2 results in the increased stability of cyclin A which in turn facilitate G1/S transition (Huang *et al*., 2011). Earlier datamining studies have shown molecular mechanisms that regulate cell cycle evolve considerably throughout animal and plants evolution (Cross *et al*., 2011; de Lichtenberg *et al*., 2007).

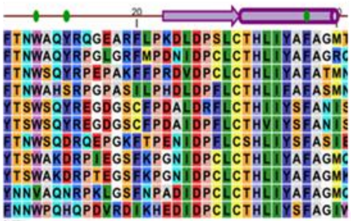
USP34 and USP9 origin is contemporaneous with the emergence of the associated molecular pathways in metazoans. USP34 deubiquitinates axin, a key molecule of Wnt/β-catenin which in turn facilitates β catenin mediated transcription (Lui *et al*., 2011). Wnt pathway is important in cell proliferation and development and first discovered in fruit fly (Baker *et al*., 1987). No homologues of Wnt have been detected in the protozoa, plants and bacteria however, 14 homologues were identified in the cnidarians suggesting the point of the origin of this pathway (Kusserow *et al*., 2005). Similar to the Wnt pathway, TGFβ pathway is also metazoan specific (Huminiecki *et al*., 2009) and human USP9X has been reported to stabilize SMAD4, one of the key modulatour of TGFβ pathway (Dupont *et al*., 2009). Considering the contemporaneous origin of Wnt and TGFβ pathways and the associated USPs suggest the possibility of co-evolution of molecular pathways and their associated USPs.

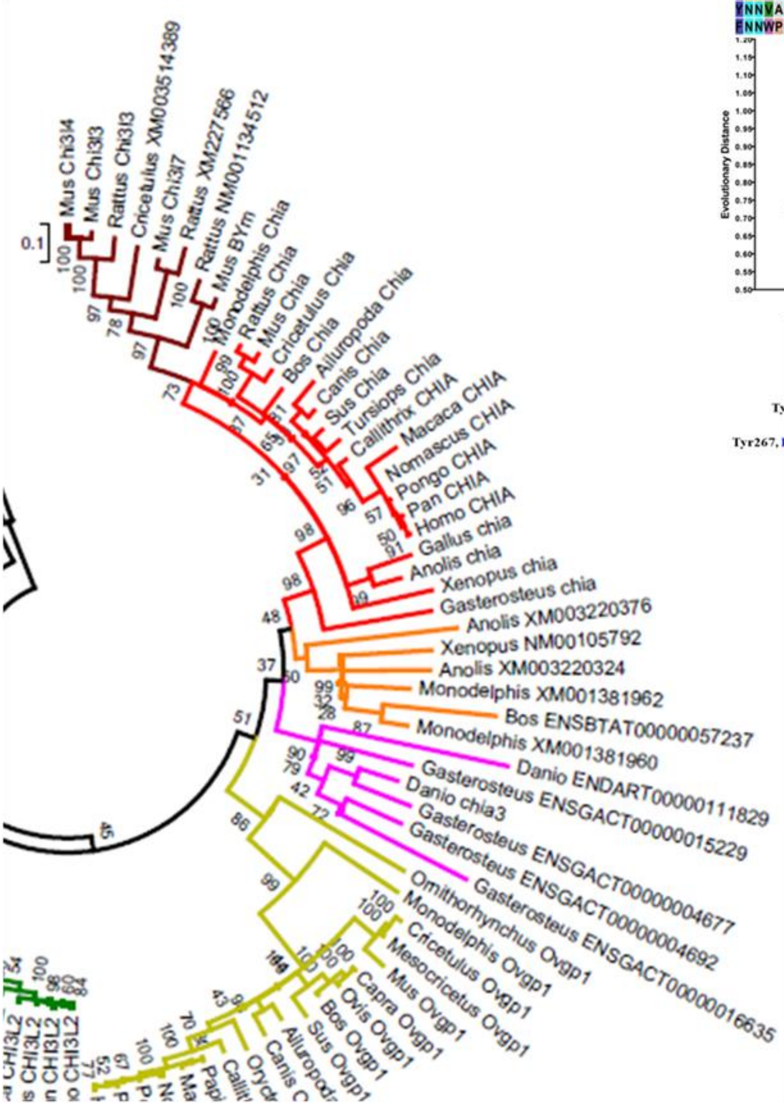
Ohnologues generated as a result of WGDs, differ in their expression profile and molecular partners (Table 5.5 and 5.6). Expression profile in humans suggests ubiquitous expression of USP2 but elevated expression of USP21 in immune related cells. Both genes are antagonist in their functions as USP2 induces cell proliferation by stabilizing MDM2 resulting in the degradation p53 (Stevenson *et al*., 2007) while USP21 inhibits the cell growth by its interaction with NEDD8 (Gong *et al*., 2000). Another such example is USP20 and USP33, USP20 expression has been observed mostly in the immune related cells while elevated expression of USP33 was observed in the nervous system. Despite the similarity in the domain organization USP20 is involved in the endocytosis whereas USP33 has been shown essential for axon guidance (Yuasa-Kawada, 2009) and centriole biogenesis (Li *et al*., 2013) potentially by its interaction with ROBO1 and centriolar protein CP110 respectively. Similarly USP4, USP11 and USP15 emergence is parallel to the emergence of vertebrates. Though expression pattern of USP4 and USP15 are similar and mostly concentrated to immune cells, USP11 expressed at high levels in different parts of nervous system.

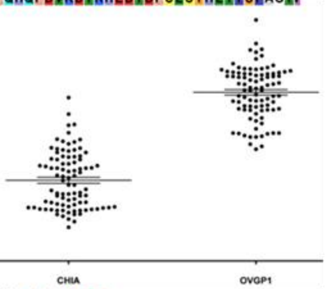
At present, functions of many USPs (especially vertebrate specific) are poorly understood or not elucidated at all. The present study highlights the important evolutionary events and mechanisms resulting in the extant array of these important proteins in vertebrates including humans. However, more insights could be gained by the structural comparison of full range of C19 domains of different USPs (once available) to resolve the ambiguous paralogous relationship. Functional innovation among different USPs though being reflected by domain composition, gene expression and molecular partners, nevertheless, studies on the selection pressure across the length of proteins may also provide useful information in this regard. Studies in this connection are underway in our research team.

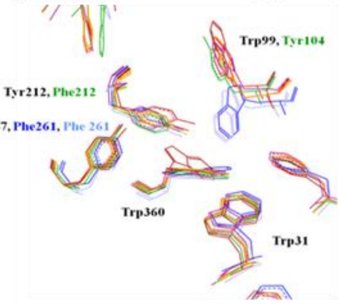


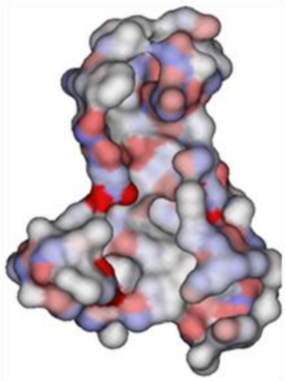
Chapter 6. Phylogenomic Studies of Chitinase and Chitinase Like Proteins


20
FTNWAQYRQGEARFLPKDLDPSLCTHLIYAFAGMT
FTNWAQYRPGLGREMPDNIDPCLCTHLIYAFAGRC
FTNWSQYRPEPAKFFPRDVDPCLCTHLIYAFATMN
FTNWAHSRPGPASILPHDLDPFLCTHLIFAFASMN
YTSWSQYREGDGSCFPDALDRFLCTHIIYSFANIS
YTSWSQYREGDGSCFPDAIDPFLCTHVIYSFANIS
FTNWSQDRQEPGKFTPENIDPFLCSHLIYSFASIE
YTSWAKDRPIEGSFKPGNIDPCLCTHLIYAFAGMC
YTSWAKDRPTEGSFKPGNIDPCLCTHLIYAFAGMK
YNNVAQNRPKLGSENPADIDPCLCTHLIYAFAGMC
ENNWPQHQPDVRDIKHEDIDPCLCTHLIYSFAGIV









6. Results: Phylogenomic studies of vertebrates chitinases and chitinase like proteins

6.1. Introduction

This chapter describes the structural phylogenomic analysis of the GH18 family homologues in vertebrates. The study includes the robust datamining of GH18 homologues in nearly all vertebrates whose genome sequence is available in public databases to explore the diversity and expansion of the genes across different vertebrate lineages. We applied the maximum likelihood method to reconstruct the phylogenetic relationship of GH18 homologues using cDNA sequences of vertebrates available on databases to explore the evolutionary relationship of vertebrate GH18 genes. Genome syntenies of vertebrates GH18 genes were also compared to investigate the underlying mechanism of GH18 gene expansion in vertebrates. The study presented here is published recently (Hussain and Wilson, 2013) and it not only helped to resolve some existing annotation issues of ChiLs but proposed novel paralogues in the vertebrate GH18 family. In addition, to explore the structural and in turn functional diversity of the proteins, protein sequence alignments and structural models of paralogous GH18 proteins were constructed and examined for the differences in functionally important residues and regions respectively.

6.2. Datamining

Two public databases, NCBI and Ensembl were extensively surveyed for GH18 family genes in mammals and other chordates. BLAST searches for GH18 family members were undertaken against the databases as well as against the genomes of the sequenced organisms. In total 388 vertebrate GH18 homologues were identified after excluding mis-annotations, duplicates and pseudogenes (Appendix VI). GH18 homologues were identified across 45 species of mammals covering 35 families and 17 orders. Homologues were found in three marsupials (opossum, Tasmanian devil and wallaby) and a monotreme (duckbill platypus), reflecting early evolutionary branches in the mammalian lineage. Relatively, fewer GH18 homologues were detected among non mammalian vertebrates such as reptile, bird, amphibian and fishes (respectively). As a representative of early vertebrates, sequences (chid1 and chitinases, respectively from sea lamprey and Arctic lamprey) from lampreys were found and included in the analysis. Digging more into the evolutionary past of vertebrates, genomic BLAST revealed one GH18 chitinase homologue in tunicate (urochordate) and two in lancelet (cephalochordate). Homologues from both sequences were incorporated in the subsequent phylogenetic analysis.

With few exceptions, both active chitinase, CHIT1 and CHIA homologues were found in most of the mammalian species examined. Additional CHIA-like homologues were identified in some species namely: cow, marmoset, bush baby, Tasmanian devil and opossum (Table.6.1). Like CHIT1 and CHIA, two ChiLs, CHIL1 and OVGP1 were also found well distributed among different mammalian lineages ranging from monotremes to primates. Consistent with the previous studies (Bussink *et al*., 2007; Srivastava *et al*., 2007), a gene identified as secretory glycoprotein 40 (BP40) was found exclusively among the members of family *bovidae* sharing 92% identity with CHIL1. Although CHIL1 is physically mapped in the cow and pig genomes, no separate locus for BP40 has been identified yet. Therefore the presence of both paralogues could be inferred as a result of species specific duplication of CHIL1 or alternatively, a single paralogue with sequencing errors or allelic variations. In contrast to CHIL1, CHIL2 was found to have a limited distribution across mammalian lineages. Except primates, homologues of CHIL2 were found to have limited distribution in different mammalian lineages. An additional array of ChiLs was observed in at least three species of family *muridae*. The mouse (as the best characterized genome) encodes Chil3, Chil4, Chil5 and Chil6. In the literature (Bussink *et al*., 2007) previously predicted mouse Chil5 and Chil6 pseudogenes have now been resolved as a single pseudogene annotated as Gm6552 in the databases. Another finding through datamining was the detection of partial CHIL2 orthologues in kangaroo rat (ENDORT00000003888), a member of rodent order which was previously thought to have lost the CHIL2 gene through evolution (Bussink *et al*., 2007). BLASTp search showed that the gene has 84% sequence identity with human CHIL2 (e value=0.0). This may suggest that the loss of CHIL2 in many rodents probably occurred after diverging from the most recent common ancestor of the order.

Chitinase genes were found in all non mammalian vertebrates examined; fishes in particular have multiple genes annotated as novel genes and showing high sequence identity with the chitinases and having catalytic motif. Surprisingly, anole lizard (reptile) does bear a ChiLs homologue. Despite showing the identical sequence and genomic location, the gene is differently annotated as CHIL1 and CHIL2 in NCBI and Ensembl databases respectively. Multiple active chitinases and ChiLs have been described among invertebrates especially arthropods (Huang *et al*., 2012). Genes annotated as oviductins are found in some arthropods, however, they showed no domain or sequence similarity with the mammalian OVGP1; rather they were found more similar (37% amino acid identity of deer tick (*Ixodes scapularis*) to human ovochymase, a serine protease. Similarly, fruit fly imaginal disc growth factor genes are also ChiLs, however, none of those genes were identified in the BLAST search using vertebrate GH18 homologues.

Species	Common name	Family	CHIT	CHIA	CHIO	OVGP	Chitinase like						CTBS	CHID
							1	2	3	4	5	6		
Mammalia														
Placental Mammals														
Afrosoricida														
Echinops telfairi	hedgehog	*Tenericidae*	■	■		■	■	■						■
Carnivora														
Ailuropoda melanoleuca	giant panda	*Ursidae*	■	■		■	■						■	■
Canis familiaris	dog	*Canidae*		■		■	■						■	■
Felis catus	cat	*Felidae*	■				■						■	■
Cetartiodactyla														
Bos taurus	cow	*Bovidae*	■	2	1	2	■	■					■	■
Bubalis bubalis	water buffalo	*Bovidae*				■	■							
Capra hircus	goat	*Bovidae*				■	■							
Ovis aries	sheep	*Bovidae*		■		■	■							■
Sus scrofa	pig	*Bovidae*	■	■		■	■	■					■	■
Tursipos truncatus	dolphin	*Delphinidae*	■	■		■	■	■					■	■
Vicugna picas	alpaca	*Camelidae*	■	■		■	■						■	
Chiroptera														
Myotis lucifugus	microbat	*Vespertilionidae*		2		■	2						■	■
Pteropus vampyrus	megabat	*Pteropodidae*				■							■	■
Euliptophyla														
Erinaceus europaeus	hedgehog	*Erinaceidae*		■		■	■							
Sorex araneus	shrew	*Soricidae*		■		■							■	■
Hyracoidea														
Procavia capensis	rock hyrax	*Procaviidae*	■			■	■	■						■
Lagomorpha														
Ochotona princeps	Am. pika	*Ochotonidae*	■			■	■	■						■
Oryctolagus cuniculus	rabbit	*Leporidae*	■			■	■	■					■	■
Perisodactyla														
Equus caballus	horse	*Equidae*	■	?		■	■	■					■	■
Primates														
Callithrix jacchus	marmoset	*Callitrichidae*	■	2		■	■	■					■	■
Gorilla gorilla	gorilla	*Hominidae*	■	■		■	■	■					■	■
Homo sapiens	human	*Hominidae*	■	■	*	■	■	■					■	■
Macaca mulatta	macaque	*Cercopithecidae*	■	■		■	■	■					■	■
Microcebus murnus	mouse lemur	*Cheirogaleidae*	■	■			■	■					■	■
Nomascus leucogenys	gibbon	*Hylobatidae*	■	■		■	■	■					■	■
Otolemur garnettii	bush baby	*Galagidae*	■	3	■		■	■					■	■
Pan troglodytes	chimpanzee	*Hominidae*	■	■		■	■	■					■	■
Papio Anubis	baboon	*Cercopithecidae*				■								
Pongo abelli	orangutan	*Hominidae*	■	■		■	■	■					■	■
Tarsius syrichta	tarsier	*Tarsiidae*	■	■			■	■					■	■
Proboscidea														
Loxodonta africana	Af. elephant	*Elephantidae*	■	■		■	■	■					■	■
Rodentia														
Cavia procellus	guinea pig	*Caviinae*	■			■	■							■
Cricetulus griseus	Ch. hamster	*Cricetidae*	■	■		■	■		■					■
Dipodomys ordii	kangaroo rat	*Heteromyidae*	■	■		■	■	?					■	■
Mesocricetus auratus	golden hamster	*Cricetidae*				■								
Mus musculus	mouse	*Muridae*	■	■		■	■		■	■	■	■	■	■
Rattus novergicus	rat	*Muridae*	■	■			■		■		■	■	■	■
Spermophilus tridecemlineatus	squirrel	*Sciruidae*	■	■		■	■						■	■
Scadentia														
Tupaia belangari	tree shrew	*Tupaiidae*		■		■	■	■					■	■
Xenarthra														
Chleopus hoffmani	two toed sloth	*Megalonychidae*	■	■		■	■	■						
Dasypus novemcintus	armadillo	*Dasypodidae*	■	■		■	■	■						
Marsupials Mammals														
Dasyuromorphia														
Sarcophilus harrissi	Tas. devil	*Dasyuridae*	■	3		■	■	■					■	■
Didelphimorphia														
Monodelphis domestica	grey opossum	*Didelphidae*	■	3	2	■	■	■					■	■
Diprotodontia														
Macropus eugenii	wallaby	*Macropodidae*	■	■		■	■	■					■	■

Cont...

Species	Common name	Family	CHIT	CHIA	CHIO	OVGP	Chitinase like						CTBS	CHID
							1	2	3	4	5	6		
Mammalia														
Egg Laying Mammals														
Monotremata														
Ornithorhyncus ananitus	platypus	*Ornithorhyncidae*	3			▨	▨						▨	▨
Aves														
Galliformes														
Gallus gallus	chicken	*Phasianidae*		3									▨	▨
Meleagris gallopova	turkey	*Phasianidae*		3									▨	▨
Passiriformes														
Taeniopygia guttata	zebra finch	*Estrilididae*		3									▨	▨
Reptilia														
Squamata														
Anolis carolinensis	anole lizard	*Poluchrotidae*	?	3	3	?		▨					▨	▨
Pelodiscus sinensis	Chinese turtle	*Trionychidae*		▨	3	?							▨	▨
Amphibia														
Anura														
Bufo japonicum	Jap.toad	*Bufonidae*			▨									
Rana catesbeiana	bullfrog	*Ranidae*		▨										
Xenopus topicalis	clawed frog	*Pipidae*		2	▨								▨	▨
Xenopus laevis	Af. frog	*Pipidae*		▨	▨								▨	▨
Pisces														
Beloniformes														
Oryzias latipes	medaka	*Adrianichthyidea*	2		3								▨	▨
Coelancanthiformes														
Latimera chalumnae	Coelacanth	*Latimeridae*		2	▨								▨	▨
Gardiformes														
Gadus morhua	cod	Gadidae		3	5								▨	▨
Gasterosterioformes														
Gasterosteus aculeatus	stickleback	*Gasterosteidae*	▨	2	4								▨	▨
Salmoniformes														
Oncorhyncus mykiss	trout	*Salmonidae*		▨	2									
Cypriniformes														
Danio rerio	zebra fish	*Cyprinidae*	▨		3									2
Perciformes														
Oreochromis niloticus			3		▨								▨	▨
Tetradontiformes														
Takifugu rubripes	fugu	*Tetradontidae*			4								▨	▨
Tetradon nigrovidis	puffer fish	*Tetradontidae*	▨		2								▨	▨
Hyperoartia														
Petromyzoniformes														
Lethenteron japonicum	Arctic lamprey	*Petromtyzontidae*	▨	▨										
Petromyzon marinus	sea lamprey	*Petromtyzontidae*												▨

Table 6.1. Distribution of GH18 homologues in vertebrates. The distribution of GH18 homologues currently identified amongst vertebrate species is shown (shaded cells). Absence or not yet identified genes are represented by blank cells.The number within the cells indicates the number of genes identified. The orthologous identification is based on the phylogenetic analysis conducted herein (described later). The newly identified (discussed later) CHIO group is also included. (?) indicates where a partial gene/unresolved has been identified.

Single genes encoding the exochitinase CTBS and ChiL CHID1 are present in most of the species examined (Table 6.1) and homologues were also observed in most organisms including plants, fungi and slime moulds.

6.3. Three evolutionary groups of vertebrate GH18 homologues

To advance the understanding of vertebrate GH18 gene family evolution, maximum likelihood trees were generated using the complete cDNA sequences of nearly all the available GH18 homologues of vertebrates and chordates. In order to root the tree, GH18 chitinase sequence of *Serratia marscens* chitinase was included. The phylogenetic tree revealed multiple clades reflecting multiple paralogous groups (Figure 6.1). Mainly, three observations could be gathered from the tree topology. First the tree shows separation of vertebrates GH18 homologues into two halves separated by the chordates GH18 homologues. On one side two clades of CHID1 and CTBS are present while on the other side both endo chitinases and associated ChiLs are present which indicates that three major evolutionary groups separated before the origin of vertebrates (Figure 6.1).

Excluding CTBS and CHID1, other expanded vertebrates GH18 homologues share lineal ancestry with GH18 homologues of non vertebrates chordates. Although both CTBS and CHID1 are retained as single gene per species (except in zebra fish where two homologues of CHID1 were noticed) the remaining GH18 homologues have expanded extensively in vertebrates especially within mammalian lineages. The limited homology (<30%) between CHID1, CTBS and the other GH18 homologues could account for instability in the alignment (especially at the phylogenetically important data points) and evidenced by low bootstrap support in the tree. Consequently, to improve the accuracy of prediction of the phylogenetic relationship of vertebrate GH18 homologues, another tree was constructed excluding CTBS and CHID1 sequences.

6.4. The expanding vertebrate GH18 homologues

Excluding CTBS and CHID1 genes, a maximum likelihood tree was reconstructed from nearly all available vertebrate GH18 homologues and rooted with the lancelet chitinase sequences (Figure 6.2). The tree reveals two major groups named here CHIT and CHIA super clades. Both super clades are further divided into smaller clades of active chitinases and ChiLs. The CHIT super clade includes one active chitinase (CHIT1) and two ChiLs: CHIL1 and CHIL2. The anole lizard ChiL is annotated in the Ensembl and NCBI databases as chi3l2 and chi3l1 respectively, clusters with the mammalian CHIL2 clade with a high bootstrap support (97%). In support to this, this ChiL shows a shorter evolutionary distance to mammalian CHIL2 homologues than to CHIL1 (Figure 6.3A). This study indicates two separate events of gene duplications, but in a different order

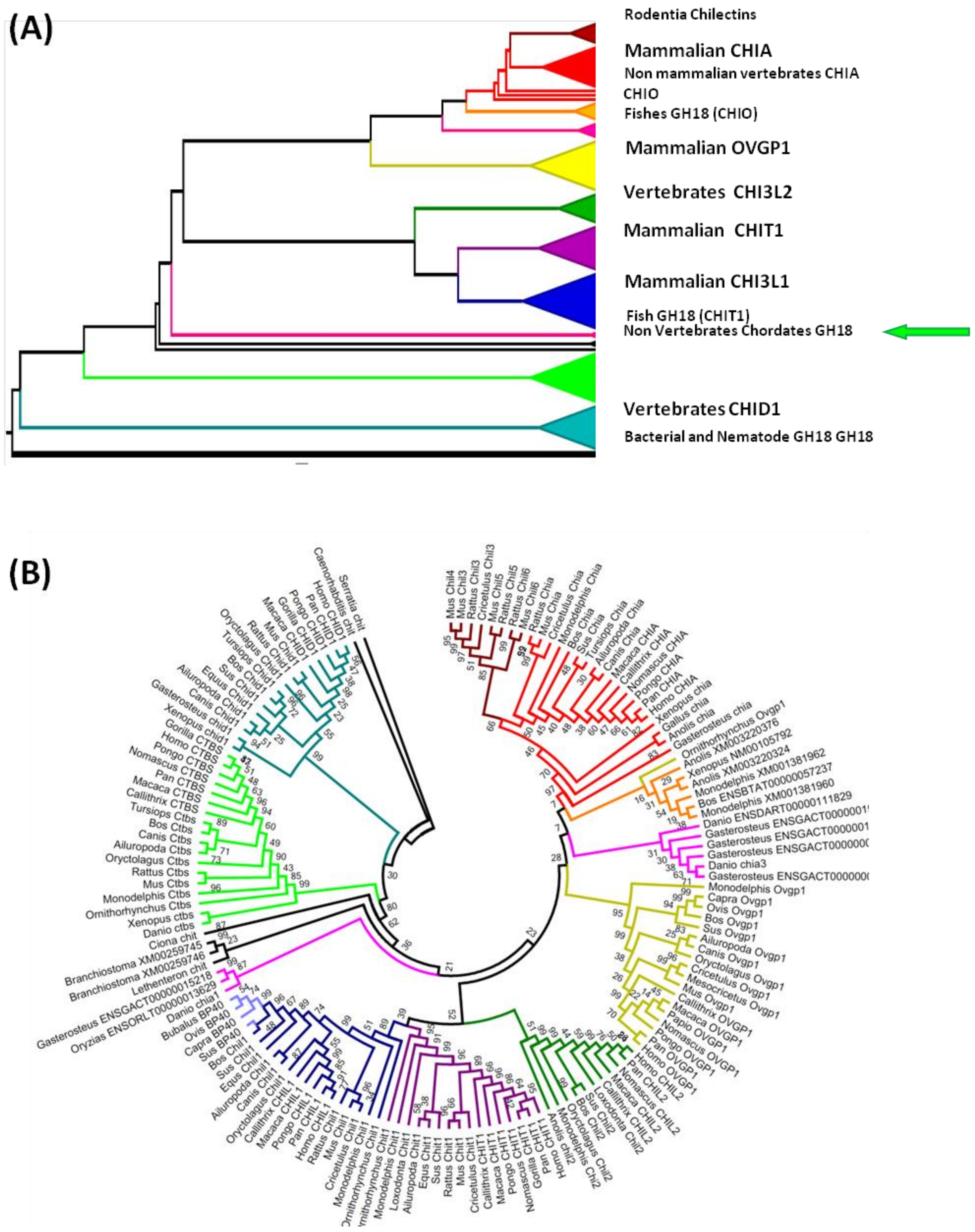


Figure 6.1. Phylogenetic tree of vertebrates GH18 homologues. Phylogenetic tree of GH18 homologues of chordates was constructed using the maximum likelihood method. The tree is shown in collapsed **(A)** and expanded format **(B)**. The tree is rooted with the *S. marcescens* chitinase homologue and support values obtained by 1000 bootstrap replicates are shown. Clades and subclades are coloured differently to represent known and potentially new paralogues. The green arrow in **(A)** indicates the branches of non vertebrate chordate GH18 homologues suggesting separation of the CTBS and CHID1 from other vertebrate GH18 homologues predates origin of the chordates.

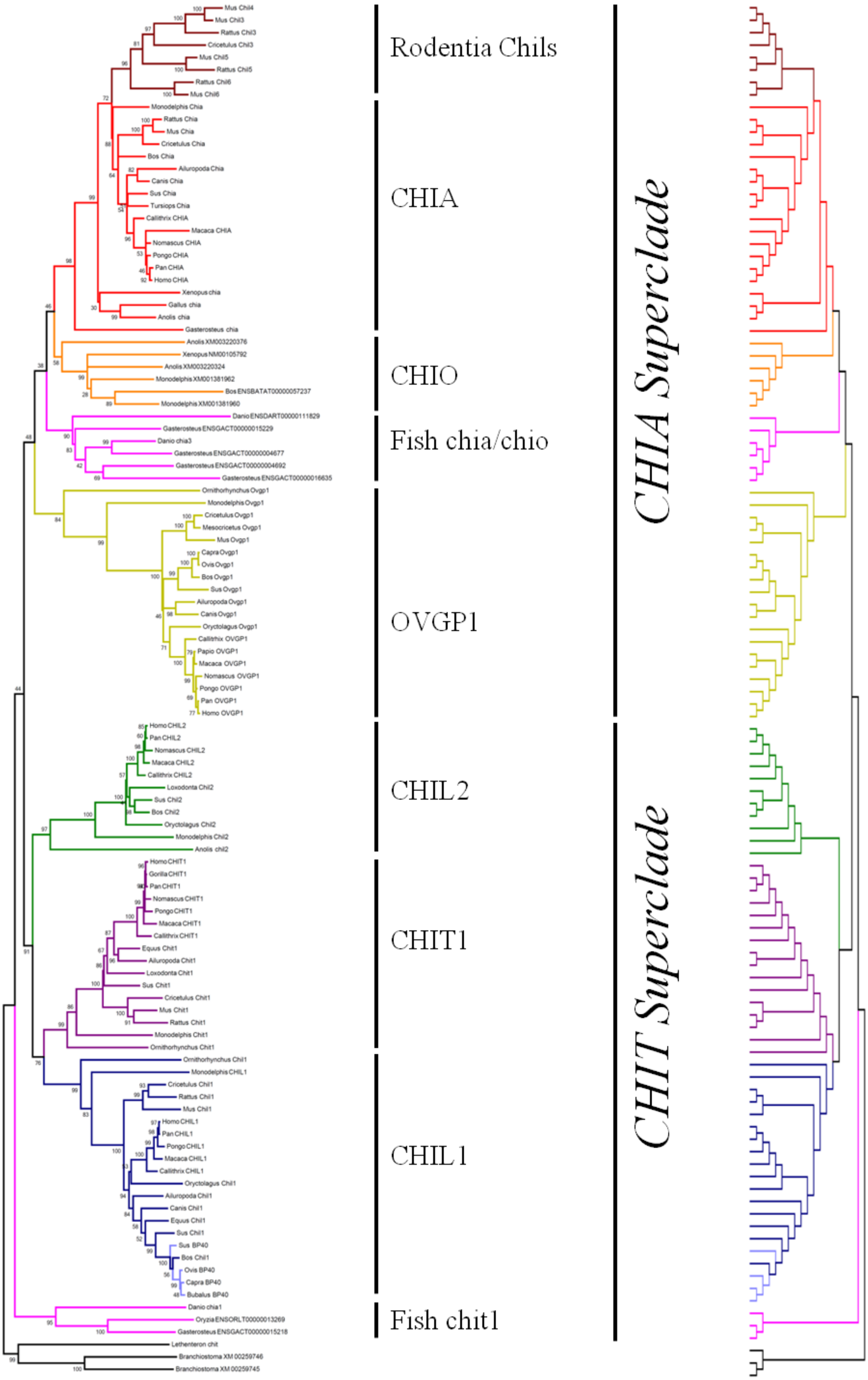

Mus Chil4
Mus Chil3
Rattus Chil3
Cricetulus Chil3
Mus Chil5
Rattus Chil5
Rattus Chil6
Mus Chil6
Monodelphis Chia
Rattus Chia
Mus Chia
Cricetulus Chia
Bos Chia
Ailuropoda Chia
Canis Chia
Sus Chia
Tursiops Chia
Callithrix CHIA
Macaca CHIA
Nomascus CHIA
Pongo CHIA
Pan CHIA
Homo CHIA
Xenopus chia
Gallus chia
Anolis chia
Gasterosteus chia
Anolis XM003220376
Xenopus NM00105792
Anolis XM003220324
Monodelphis XM001381962
Bos ENSBATAT00000057237
Monodelphis XM001381960
Danio ENSDART00000111829
Gasterosteus ENSGACT00000015229
Danio chia3
Gasterosteus ENSGACT00000004677
Gasterosteus ENSGACT00000004692
Gasterosteus ENSGACT00000016635
Ornithorhynchus Ovgp1
Monodelphis Ovgp1
Cricetulus Ovgp1
Mesocricetus Ovgp1
Mus Ovgp1
Capra Ovgp1
Ovis Ovgp1
Bos Ovgp1
Sus Ovgp1
Ailuropoda Ovgp1
Canis Ovgp1
Oryctolagus Ovgp1
Callithrix OVGP1
Papio OVGP1
Macaca OVGP1
Nomascus OVGP1
Pongo OVGP1
Pan OVGP1
Homo OVGP1
Homo CHIL2
Pan CHIL2
Nomascus CHIL2
Macaca CHIL2
Callithrix CHIL2
Loxodonta Chil2
Sus Chil2
Bos Chil2
Oryctolagus Chil2
Monodelphis Chil2
Anolis chil2
Homo CHIT1
Gorilla CHIT1
Pan CHIT1
Nomascus CHIT1
Pongo CHIT1
Macaca CHIT1
Callithrix CHIT1
Equus Chit1
Ailuropoda Chit1
Loxodonta Chit1
Sus Chit1
Cricetulus Chit1
Mus Chit1
Rattus Chit1
Monodelphis Chit1
Ornithorhynchus Chit1
Ornithorhynchus Chil1
Monodelphis CHIL1
Cricetulus Chil1
Rattus Chil1
Mus Chil1
Homo CHIL1
Pan CHIL1
Pongo CHIL1
Macaca CHIL1
Callithrix CHIL1
Oryctolagus Chil1
Ailuropoda Chil1
Canis Chil1
Equus Chil1
Sus Chil1
Sus BP40
Bos Chil1
Ovis BP40
Capra BP40
Bubalus BP40
Danio chia1
Oryzia ENSORLT00000013269
Gasterosteus ENSGACT00000015218
Lethenteron chit
Branchiostoma XM 00259746
Branchiostoma XM 00259745
Rodentia Chils
CHIA
CHIO
Fish chia/chio
OVGP1
CHIL2
CHIT1
CHIL1
Fish chit1
CHIA Superclade
CHIT Superclade

Figure 6.2. Phylogenetic tree of expanding vertebrate GH18 homologues. Phylogenetic tree of GH18 homologues (excluding CTBS and CHID1) of chordates was reconstructed using the maximum likelihood method employing the General Time Reversible (GTR) model with 1000 boot strap replicates. The tree is rooted with *B. floridae* GH18 homologues. Clades and subclades are coloured differently to represent known and potentially new paralogues. The tree is shown in both branch length (left) and topology format (right). The clades at the root are generally supported with high boot strap values (>90%). Note the two clades of fish GH18 homologues (pink) and new paralogous clade (CHIO).

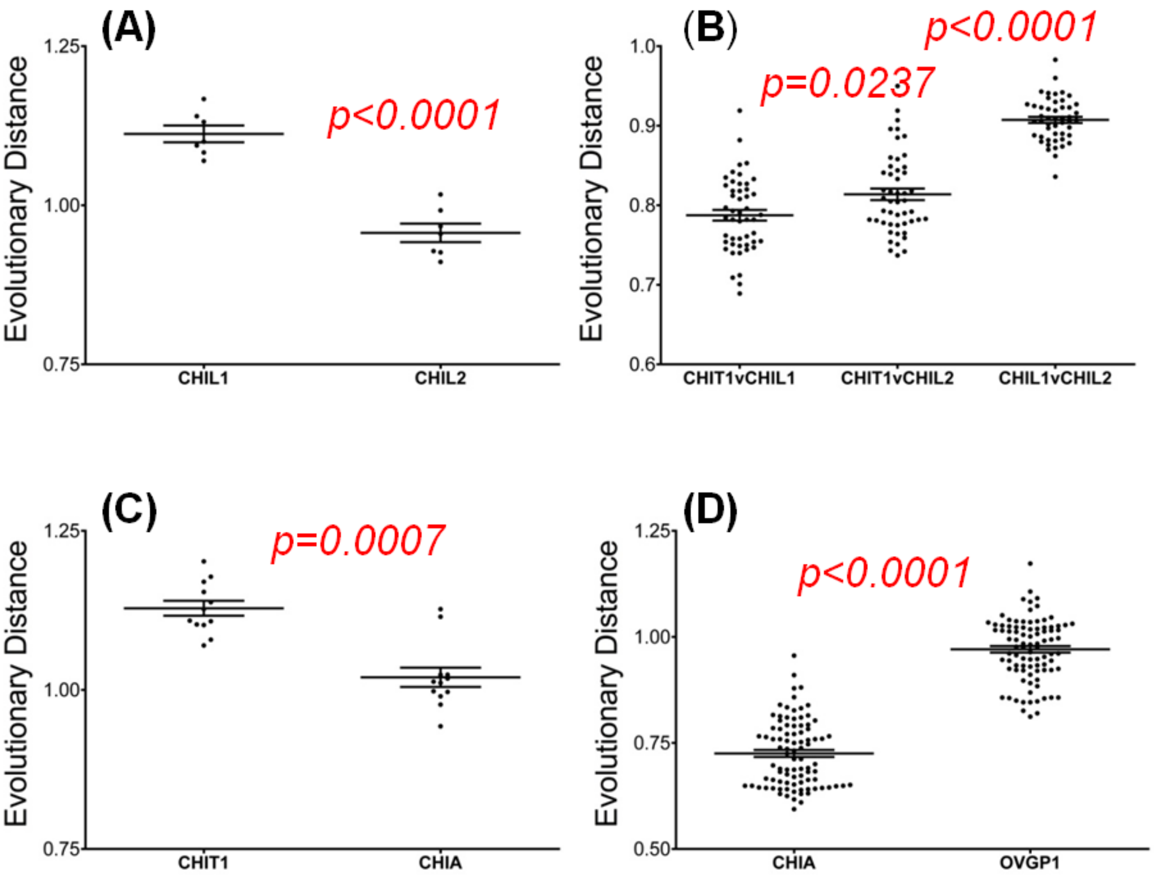


Figure 6.3. Evolutionary distance analysis of selected GH18 homologues. Evolutionary distances of selected GH18 homologues were estimated using the maximum composite likelihood method. The data distribution and statistical significance were tested using the Kolmogorov-Smirnov and the Wilcoxon tests respectively (*p* values are shown in red). Error bars indicate standard error of mean (SEM). **(A)** Anolis lizard chil2 evolutionary distance is compared with CHIL1 and CHIL2 homologues of selected mammalian species (human, chimpanzee, marmoset, rhesus monkey, pig, cow and opossum). **(B)** Evolutionary distance of homologues of CHIT1, CHIL1 and CHIL2 from the same species were compared in pairs (as indicated). **(C)** Evolutionary distance of clawed toad NM00105790 (chit1) was compared against CHIT1 and CHIA homologues of human, baboon, orangutan, rhesus monkey, pig, panda, rat, hamsters and opossum. **(D)** Evolutionary distance of CHIO homologues were compared against all homologues (included in Figure 6.2) of CHIA and OVGP1.

(Bussink *et al*., 2007; Funkhouser and Aronson, 2007) leading to the emergence of ChiLs where CHIL2 origin predating the CHIL1. To support this notion, evolutionary distances of comparable orthologues of CHIT1, CHIL1 and CHIL2 were estimated (Fig.6.3B). This comparison indicates that the evolutionary distance between CHIT1 and CHIL1 is significantly smaller in comparison to CHIT1 and CHIL2. Moreover, the distance between CHIL2 and CHIL1 is significantly (p<0.0001) larger than CHIL1 and CHIT1. Taken together these data indicate that the CHIL2 ancestor emerged prior to the reptilian-mammals split and CHIL1 emerged later from a gene duplication of ancestral CHIT1 before the diversification of mammals.

Unlike the CHIT1 clade, which is comprised of only mammalian homologues, the CHIA clade, contains one representative of reptile, bird, amphibian and fish with 98% bootstrap support. This would suggest that there has been an extensive gene loss of CHIT1 homologues in the non mammalian vertebrates. Like the CHIT super clade, the CHIA super clade also includes ChiLs: OVGP1 and the rodentia specific Chils. These ChiLs form separate clades with high bootstrap support, 97% for rodentia specific chilectins and 86% for OVGP1 (99% excluding platypus Ovgp1). Rodentia Chils form a separate cluster within the CHIA clade which points to a recent evolutionary origin from rodentia Chia. The gene currently annotated as rat Chi3l4 clusters with the mouse Chil5 (100% bootstrap support) suggesting an orthologous relationship between these two homologues. Between the CHIA and OVGP1 clades two relatively sparsely populated clades are present referred to as CHIO and fish chio/chia herein (Figure 6.2). The CHIO clade contains representative of genes from a reptile, an amphibian and mammals (opossum and cow) and is statistically (99%) supported as monophyletic excluding one (XM003220376) of the two Anolis homologues within this clade. The CHIO clade also contains a clawed toad homologue (NM001056792) which is annotated as chit1 in both NCBI and Ensembl databases. This misnomer has led to the impression in earlier studies (Bussink *et al*., 2007) that duplication leading to CHIA and CHIT1 occurred at the common ancestor of tetrapods. However, phylogenetic reconstruction based on increased taxa does not support this idea as clawed toad “chit1” (NM001056792) does not cluster with the CHIT1 clade or CHIT super clade, rather it joins the separate monophyletic CHIO clade within the CHIA superclade. Furthermore, evolutionary distance analysis also showed the same homologue of clawed toad is statistically closer (p=0.0024) to CHIA as compared to CHIT1 (Figure 6.3C). Although CHIO representatives form a distinct clade in between CHIA and OVGP1 but the presence of DXDXE motif suggests they are potentially active chitinases. Evolutionary distance analysis also indicates that CHIO genes are more closely related (p<0.001) to CHIA than OVGP1 (Figure 6.3D). Another smaller clade, in between OVGP1 and CHIA is of fishes GH18 homologues (90% bootstrap support) which are annotated as CHIA-like or novel gene in NCBI and ENSEMBL respectively. The presence of a DXDXE motif in all the representatives (except stickle back ENSGACT00000016635) suggest they are active chitinases. Three fish and one lamprey GH18 homologues with two discrete branches are present at the base of both super clades which raises two possibilities, one that these fish genes (at least those of bony fishes) are CHIT1 orthologues but cannot yet be reliably placed because of the weak phylogenetic signals. Alternatively, it is possible that these homologues are the lineal common ancestor of all the vertebrate CHIT1 and CHIA genes as suggested in earlier studies (Funkhouser and Aronson, 2007; Huang *et al.*, 2012). However, the presence of bony fish GH18 homologues in the CHIA super clade and more specifically CHIA clade support the former inference.

To explore if outgroup (lancelet chitinases) biases on the tree topology two separate phylogenetic trees were reconstructed excluding lancelet and lamprey genes (Figure 6.4A) and rooted with a more ancestral GH18 homologue of nematode worm and including all the chordates GH18 homologues (Figure 6.4B). The fish homologues change location in the tree demonstrating a lack of resolution in determining their evolutionary history. Despite this the overall topology of the tree remains intact with high boot strap support, indicating the stability of the tree. In the tree devoid of both lancelet and lamprey chitinases, fish GH18 homologues (with the notable exception of stickleback chia) are present at the base of both CHIA and CHIT super clades forming two distinct clades of fish chio/chia and fish chit1. Whereas in the tree rooted with the nematode worm homologue, the fish chio/chia clade is present at the base of CHIA superclade and fish chit1 clade is situated at the base of both CHIT and CHIA superclades.

Considering the data of all trees together it appears that CHIL1 and CHIL2 arose from separate duplication events. In addition a new paralogous group, CHIO has been identified.

6.5. Novel paralogues of vertebrates GH18 family

Data mining has shown multiple CHIA like genes in several mammals and non mammalian vertebrates. Some of these genes form a discrete clade, resolving between CHIA and OVGP1 clades and referred here as CHIO. To explore these sequences further, separate phylogenetic trees were reconstructed including all CHIA like genes selected OVGP1 orthologues with lamprey chitinase and lancelet chitinases to root the tree (Figure 6.5). The tree shows one major clade of CHIA from orthologues ranging from fishes to mammals with strong bootstrap support (99%). However, this major clade is intervened by at least two other groups, which reflect the known speciation events (Bininda-Emonds *et al.*, 2007). For example a new clade referred to as CHIAII contain two primate

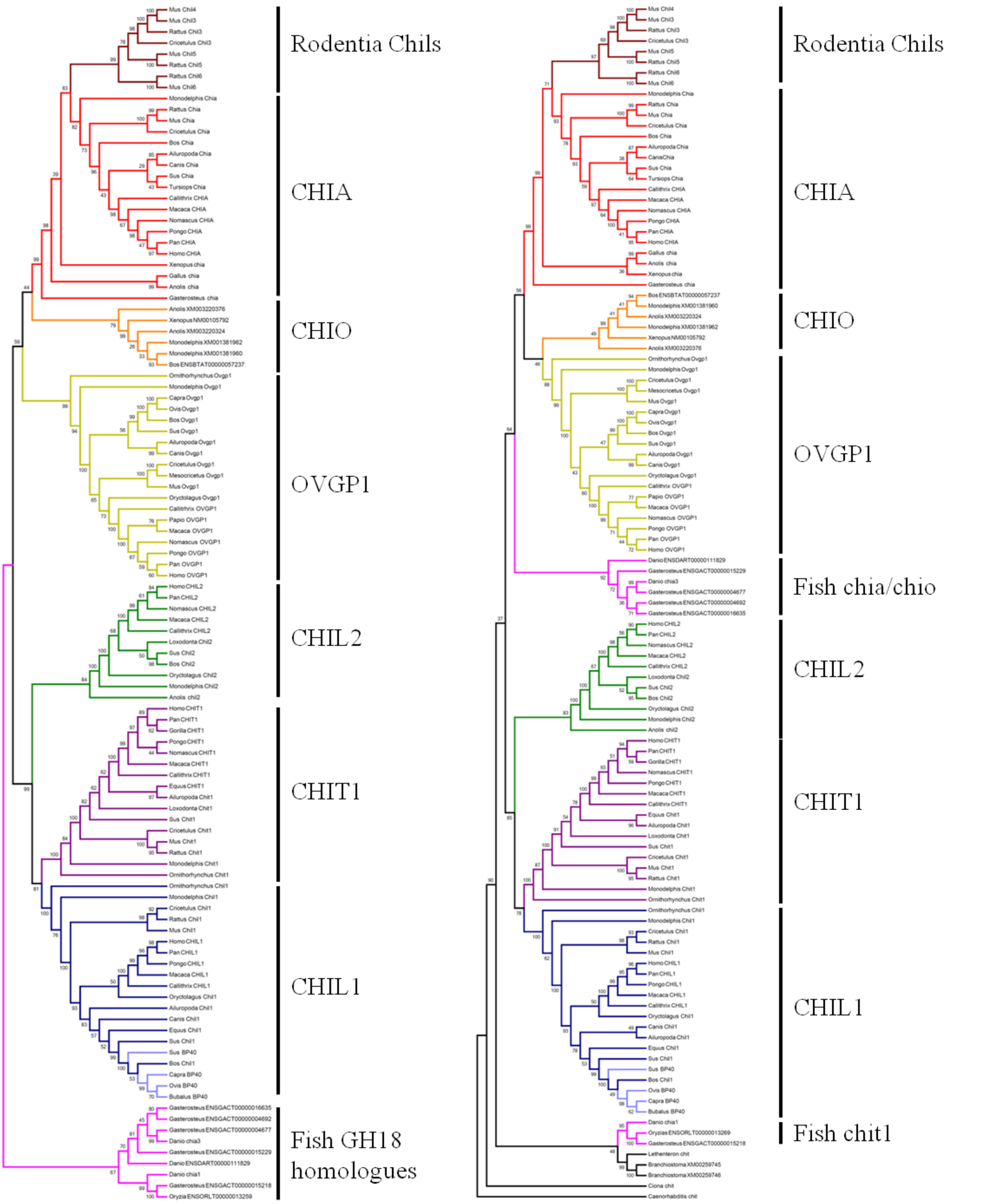


Figure 6.4. Phylogenetic relationship of vertebrate GH18 homologues. Two different phylogenetic trees were reconstructed using the maximum likelihood method and employing GTR model of nucleotide substitution by excluding lancelet and lamprey **(A)** and including tunicate and nematode worm GH18 homologues **(B)** in the taxa included in the Fig.6.2. The Clades are coloured differently to represent the paralogous relationship. Note the swapping of relative position in the trees of the fish GH18 compared to the tree shown in Fig. 6.2. Also note, stickle back chia remains in the CHIA clade in both trees.

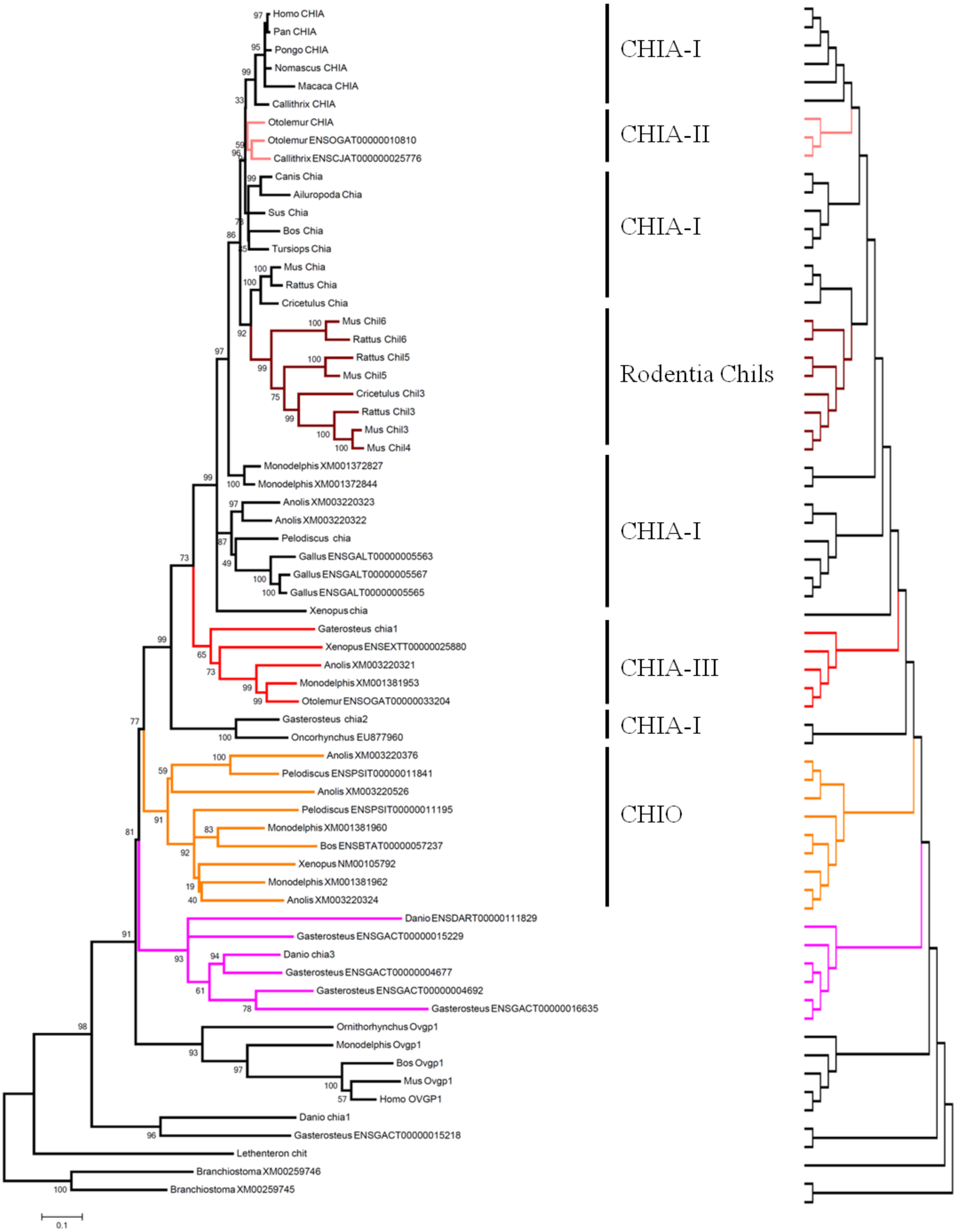


Figure 6.5. Evolutionary tree of vertebrate CHIA super clade genes. The evolutionary history of the CHIA superclade was reconstructed using all available complete mammalian CHIA related genes and selected OVGP1 homologues as well sequences from representatives of fish species. The tree was reconstructed using the maximum likelihood method using the GTR model of nucleotide substitution with 1000 bootstrap replicates and rooted with *B. floridae* chitinase genes. Different paralogous clades are labelled including three newly identified paralogues i.e. CHIA-II, CHIA-III and CHIO.

homologues (bush baby and marmoset), and is located between the Chia of carnivores and primates. Similarly, a CHIAIII clade includes homologues from fish to mammals, located in the tree between clawed toad and stickleback chia. In addition to this the rodentia specific Chils clearly show common ancestry with rodentia Chia suggesting their recent origin. All CHIA clades collectively share a common ancestry with the CHIO clade (91% bootstrap support) which is populated by three mammalian homologues (2 of opossum and 1 of cow), one amphibian sequence clawed toad (NM00105792; database annotation chit1) and five reptilian sequences (3 of anole lizard and 2 of Chinese soft shell turtle). Finally, both CHIO and CHIA clades share common ancestry with the other GH18 homologues of fishes. Both OVGP1 and potential fish chit1 form separate and distinct clades from CHIAs and CHIO. In addition to the newly identified paralogous clades (CHIAII, CHIAIII and CHIO), some additional species specific duplications in chia were observed in opossum, anole lizard and chicken (Figure 6.5). Further genomic BLAST search also revealed the presence of three pseudogenes in humans. To investigate the evolutionary relationship of these pseudogenes, the CHIA phylogenetic tree was reconstructed incorporating the human pseudogenes (Figure 6.6). The tree suggests that relics of all three novel paralogues (CHIA-II, CHIA-III and CHIO) are present in the human as pseudogenes, providing evidence of extensive gene death of GH18 members across the mammalian lineage. It might be expected that pseudogenes would show long branches because of higher nucleotide substitution rates due to possible loss of selection pressure. However, these 3 human pseudogenes do not show long branches and do not greatly distort the tree topology.

Collectively, the data demonstrate that the CHIA ancestral genes have undergone extensive gene duplication events followed by extensive gene death in many examined mammalian lineages during the evolutionary course. At least two duplication events could be aligned before the emergence of mammals giving rise to CHIO/OVGP1 and CHIA-III. Another duplication event may have occurred in the common ancestor of primates giving birth to CHIA-II. Consistent with the birth and death model of gene evolution, three human pseudogenes (and three from macaque, Appendix VI) reflect death of genes in each of the three newly identified paralogous groups.

6.6. Fish paralogues of vertebrate GH18

To examine the proximal time line for the birth and diversification of CHIT1/CHIA and CHIA/CHIO clades, fish sequences were investigated in more detail. Therefore, a phylogeny of fish GH18 homologues was reconstructed using the maximum likelihood method including two mammalian CHIA (Figure 6.7). The tree composed of 4 distinct

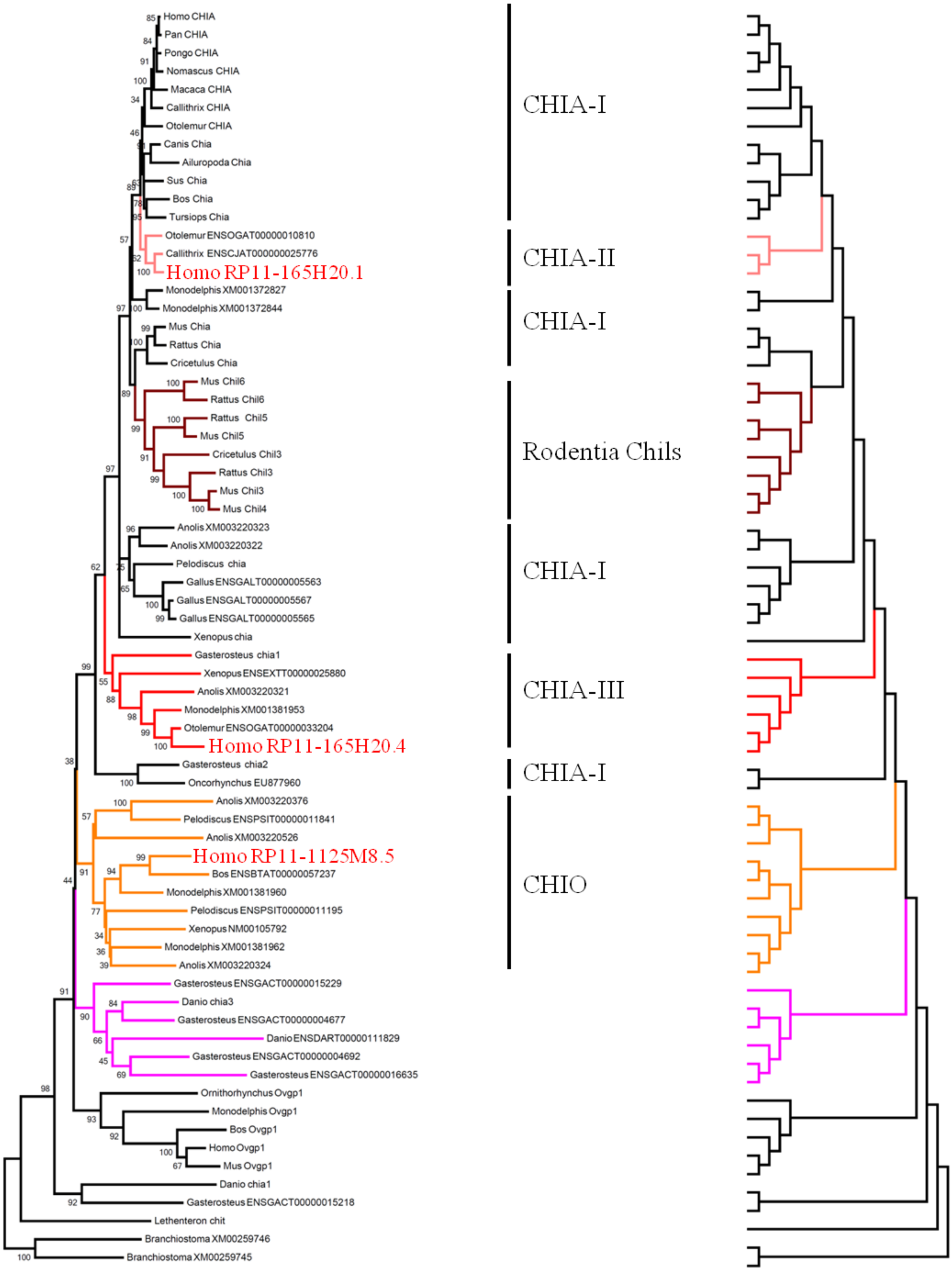


Figure 6.6. Phylogenetic relationship of human pseudogenes with CHIA superclade genes. The evolutionary tree of CHIA super-clade homologues was reconstructed using all available complete mammalian CHIA related genes, human pseudogenes and selected OVGP1 homologues as well sequences from representatives of fishes. The tree was reconstructed using the maximum likelihood method employing the GTR model of nucleotide substitution with 1000 bootstrap replicates and rooted with lancelet chitinase genes. Different paralogues clades are labelled including three newly identified paralogues i.e. CHIA-II, CHIA-III and CHIO. Note the placement of human pseudo genes in all three of the newly identified paralogous clades.

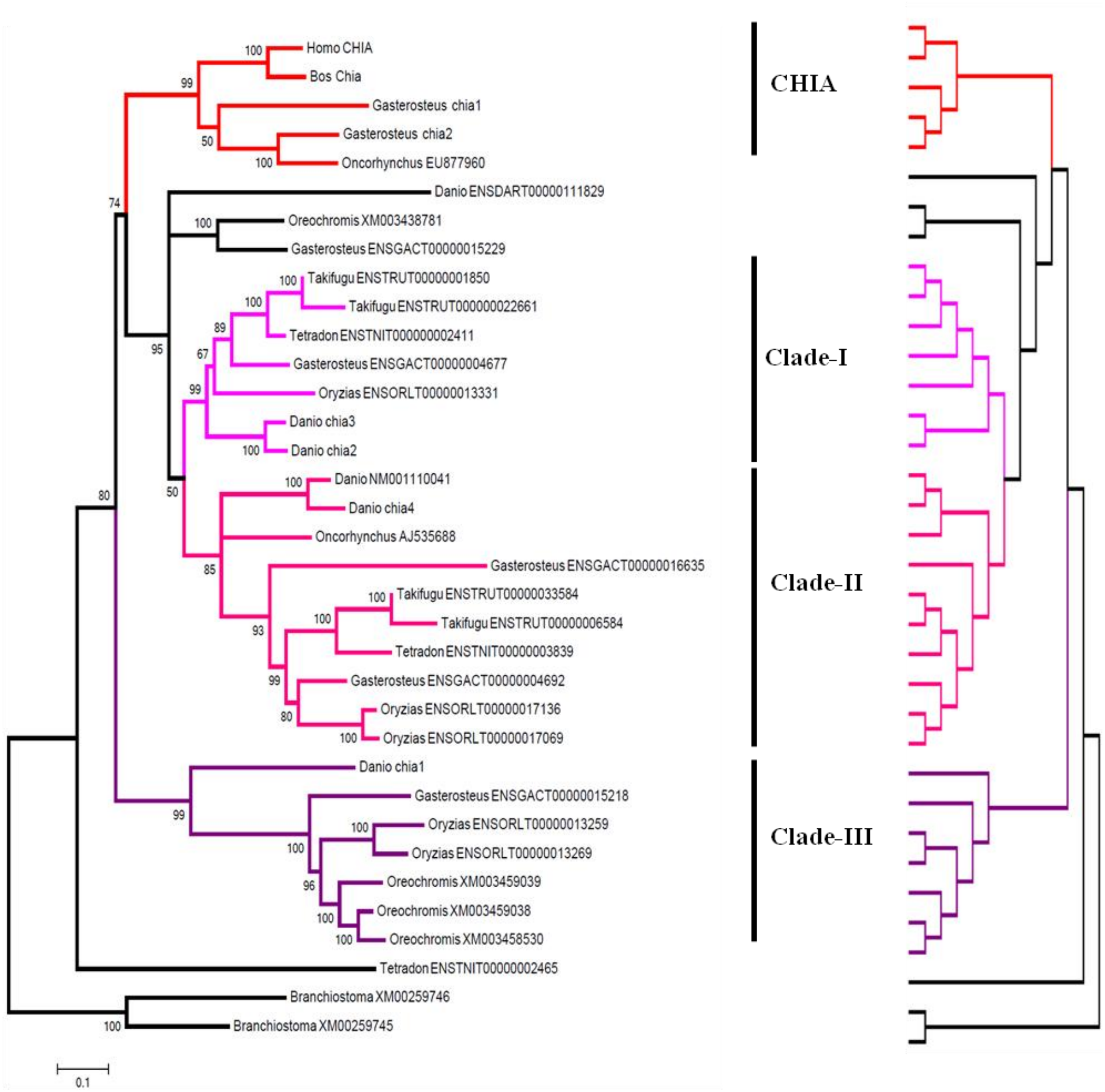

100
Homo CHIA
99
Bos Chia
Gasterosteus chia1
50
Gasterosteus chia2
100
Oncorhynchus EU877960
74
Danio ENSDART00000111829
100
Oreochromis XM003438781
Gasterosteus ENSGACT00000015229
100
Takifugu ENSTRUT00000001850
100
Takifugu ENSTRUT000000022661
89
Tetradon ENSTNIT000000002411
95
67
Gasterosteus ENSGACT00000004677
99
Oryzias ENSORLT00000013331
Danio chia3
100
Danio chia2
100
Danio NM001110041
80
50
Danio chia4
Oncorhynchus AJ535688
85
Gasterosteus ENSGACT00000016635
100
Takifugu ENSTRUT00000033584
100
Takifugu ENSTRUT00000006584
93
Tetradon ENSTNIT00000003839
99
Gasterosteus ENSGACT00000004692
80
Oryzias ENSORLT00000017136
100
Oryzias ENSORLT00000017069
Danio chia1
Gasterosteus ENSGACT00000015218
99
100
Oryzias ENSORLT00000013259
100
Oryzias ENSORLT00000013269
96
Oreochromis XM003459039
100
Oreochromis XM003459038
100
Oreochromis XM003458530
Tetradon ENSTNIT00000002465
Branchiostoma XM00259746
100
Branchiostoma XM00259745
0.1
CHIA
Clade-I
Clade-II
Clade-III

Figure 6.7. Evolutionary relationship Fish GH18 homologues. The phylogenetic tree of fish GH18 homologues (excluding CTBS and CHID1) was reconstructed using the maximum likelihood method using the GTR model with 1000 bootstrap replicates. Four distinct phyletic clades are coloured differently namely CHIA and clade I-III. Clade I and II are potential homologues of CHIO while homologues in clade III are potentially the fish representatives of CHIT1.

phyletic groups. First acting as an outgroup of the CHIA clades is a clade of 7 genes (clade-III; 99% bootstrap support) belonging to GH18 homologues of medaka, zebrafish, Nile tilapia and stickle back. This separation is consistent with the hypothesis that these genes may be CHIT1 orthologues. Three fish sequences, two from stickleback and one from trout, formed a clade with mammalian CHIA (CHIA clade; 99% bootstrap support) strongly supporting the orthologous relationship of these homologues with the mammalian CHIA. Most of the remaining fish GH18 homologues occupied two distinct phyletic groups (clade-I and clade-II) between the CHIA clade and clade-III, with evidence of further duplications in several species. Four fish homologues did not cluster with any of designated clades namely: stickleback ENSGACT00000015229, medaka XM00343878, zebrafish ENSDRT00000111829 (forming outgroup of CladeI and Clade II) and pufferfish ENSTNIT00000002465.

Consistent with all the data is that the duplication event giving rise to the ancestral CHIT1 and CHIA, and probably CHIO as well, occurred prior to the divergence of bony fishes.

6.6. Genomic synteny reflects the phylogeny

The phylogenetic relationship of chitinases and ChiLs is recapitulated in the chromosomal location and gene order of these genes (Figure 6.8). The distantly related CTBS and CHID1 genes are not chromosomally linked to other GH18 homologues in any of the vertebrate species examined. The Chia related ChiLs, OVGP1 and rodentia specific Chils are located in the proximity of the CHIA genes in the analysed organisms. Similarly, CHIL1 and CHIT1 are consistently present next to each other. Interestingly, CHIL2 which showed a close evolutionary relationship with CHIT1 in the phylogenetic analysis was found genomically linked with CHIA. This observation points to the birth of this paralogue predating the genetic rearrangement event that physically separated the CHIT1 and CHIA loci, while CHIL1 could have arisen subsequent to this separation or been carried with CHIT1. Interestingly, these genes are all present in close proximity within same chromosome in the anole lizard. This points to the physical separation of CHIT1 and CHIA genes seen in mammals, post dating reptilian mammalian split. In fishes, though multiple GH18 homologues were found, many of them have not been fully mapped on their respective genomes. However, the homologues in fishes are generally distributed on two different chromosomes which could suggest that they are the product of fish specific whole genome duplication. Moreover, in stickle back and zebra fish, homologues associated with three different phyletic clades (CHIA, cladeI/II and clade III) have been physically mapped on the same chromosomes, which indicate that at least two of gene duplication events may

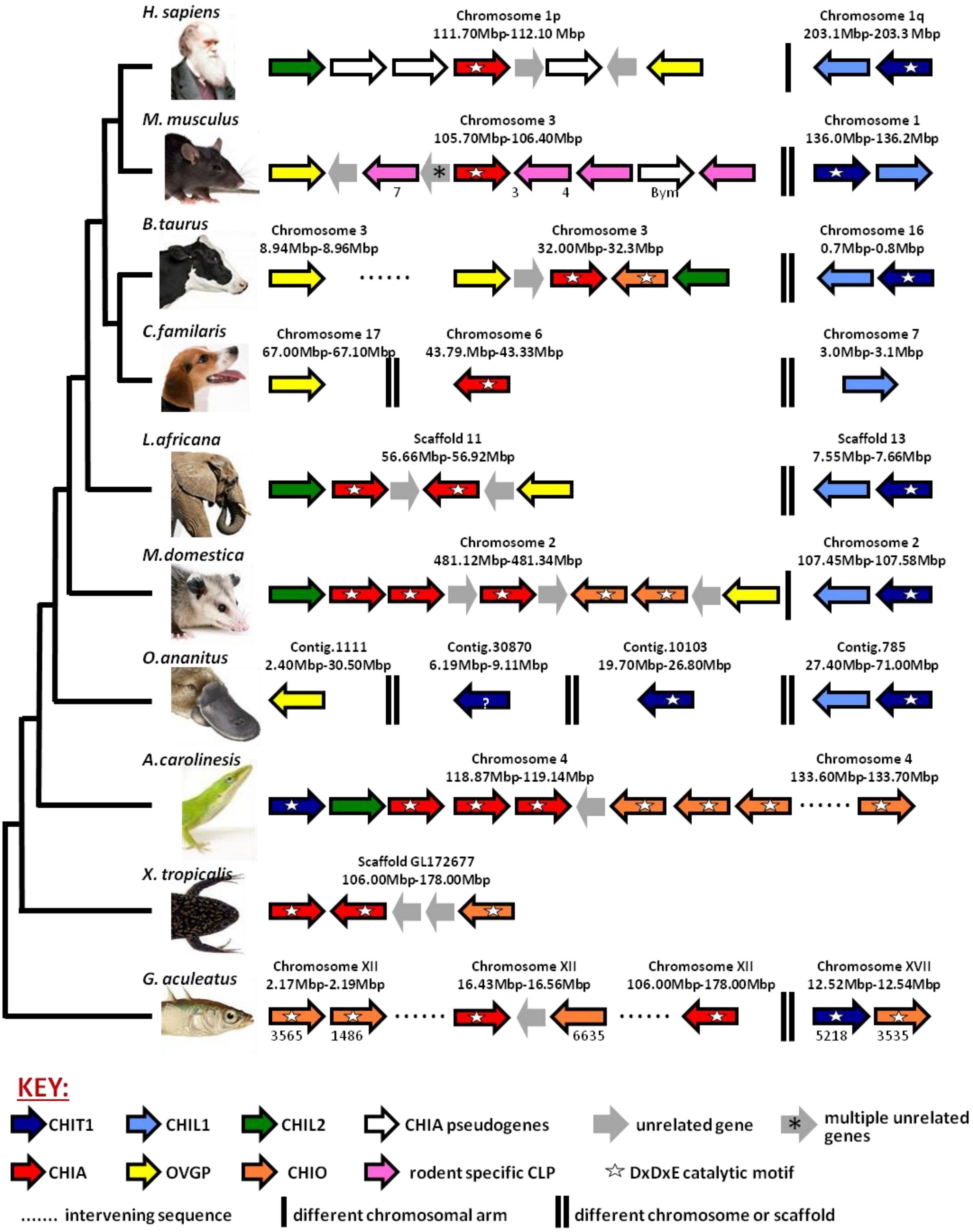


Figure 6.8. Genomic synteny of vertebrate GH18 homologous. The chromosomal locations of Chi/ChiL genes of representative vertebrate species are depicted. Arrow heads correspond to the direction of transcription. Genes are coloured in relation to phylogenetic clades. Human pseudogenes, as shown from left to right are RP11-165H20.4, RP11-165H20.1 and RP5-1125M8.5 and in mouse Gm6522. Where genes are not annotated the last four numbers of the accession number have been used.

have occurred prior to the divergence of fishes. It is interesting to note that among fishes, the GH18 homologues are flanked with genes that are homologous of those flanking both CHIT1 and CHIA in human. For instance, one side of the predicted CHIT1 orthologoues (ENSORLT00000013259 on chromosome 5) of medaka, genes lie *CNTN2* and *NAFSc* which were found in the proximity of human CHIT1. On the other side of the fish genes lie *CDC40* and *SLC2A1*, homologous to those seems at the human CHIA locus. This observation supports the idea that the physical separation of the CHIT1 and CHIA loci seen in mammals had not occurred prior to the divergence of fish.

On the basis of the cumulative phylogenomic analyses, it seems likely that the ancestral chitinase gene underwent two duplication events leading to the formation of CHIA, CHIO and CHIT1 prior to the divergence of bony fish. Most of the further gene duplications have occurred in the recent common ancestral lineage of all mammals or some specific families (rodentia and bovidae) leading to the formation of an array of ChiLs. However, one such event of duplication in ancestral CHIT1 has occurred prior to the origin of mammals (in reptiles) resulting in the birth of CHIL2.

6.7. The tale of the OVGP1 tail

While OVGP1 orthologues show common ancestry with CHIA and CHIO and the gene has the GH18 homology domain, the C-terminal region of mammalian OVGP1 is quite different from mammalian CHIA C-terminal chitin binding domain (CBM14). In fact the extended C-terminal region of OVGP1 shares a patchy sequence similarity with mucin like proteins and previously reported as heavily glycosylated (Buhi *et al*., 2002). N and O linked glycosylation sites on proteins were predicted using CBS NetNglyc and NetOglyc servers respectively. The threshold value has been set at 0.5 at which 96% accuracy is expected, however some sites may be missed. Asn is the target residue for N-linked glycosylation with Asn-X-Ser/Thru as a consensus site, whereas O-linked glycosylation occurs on Ser or Thru residues in a context dependent manner. *In silico* prediction showed that most mammalian OVGP1 sequences with an extended C-terminal tail have multiple glycosylation sites in comparison to CHIA (Fig.6.9). Extensive glycosylation could contribute to the increased stability and viscosity of the protein and could also provide additional ligand binding sites. The OVGP1 C-terminal region showed considerable variation across mammalian species in terms of length and glycosylation sites, suggesting that variation/retention in the number of glycosylation sties may have acted as a selection force in the evolution of OVGP1 orthologues. No distinct OVGP1 orthologue was identified in the non mammalian vertebrates, however CHIO orthologues of anole lizard, clawed toad and stickle back, despite having the catalytic motif (DXDXE), have

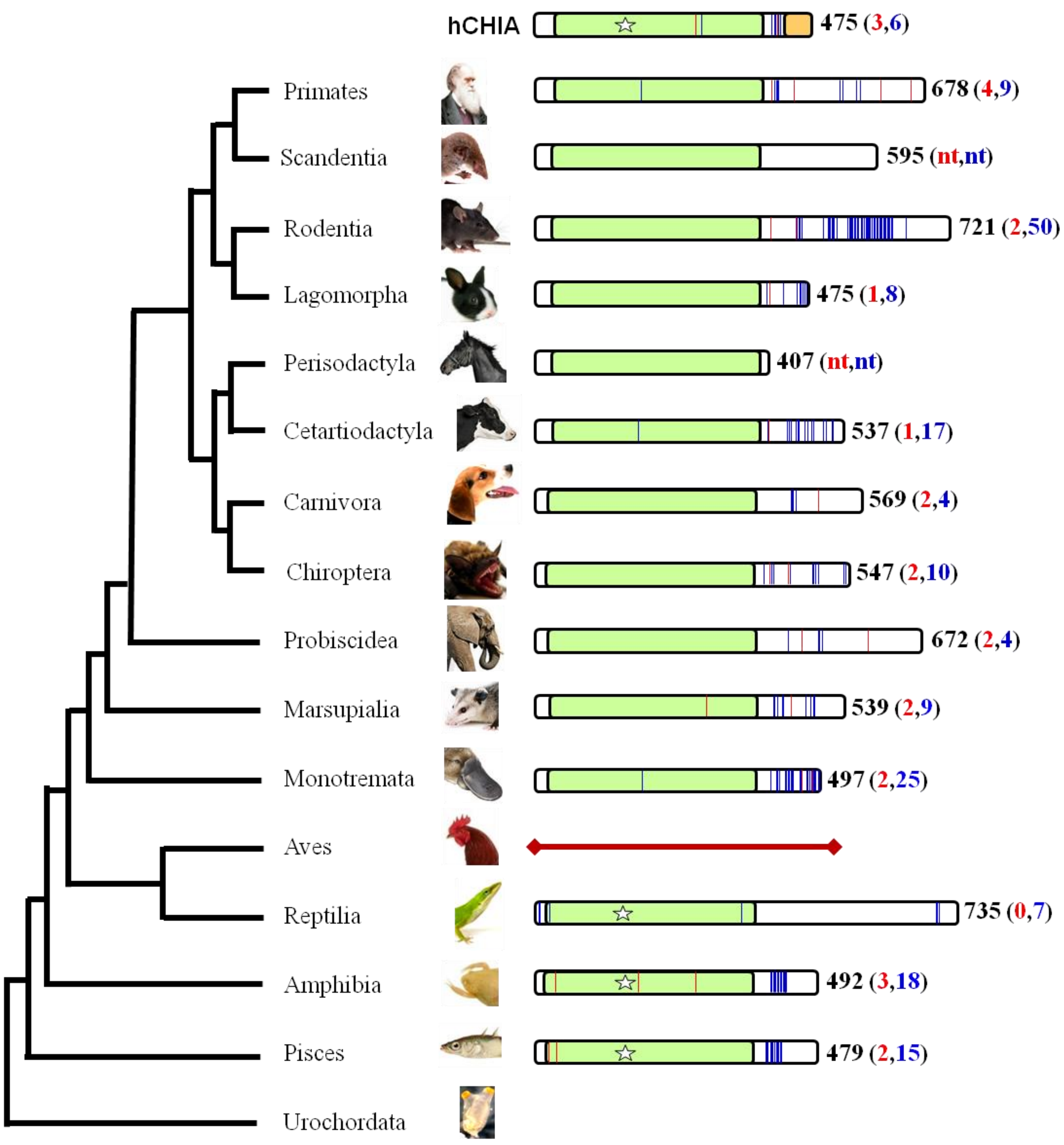


Figure 6.9. OVGP1 glycosylation. Schematic representation of OVGP1 or CHIO (where OVGP1 is not found) from several species of different vertebrates are shown. The length of the horizontal rectangle is scaled corresponding to the polypeptide length (indicated at the right of each). Human CHIA is shown at the top with GH18 domain highlighted in light green and CBM14 domain in yellow. N and O glycosylation sites (predicted using CBS servers NetNglyc and NetOglyc respectively) are shown with red and blue bars respectively. The presence of the catalytic motif is represented by a star. The sequences of horse (*E.caballus*) and shrew (*T.bellangari*) have gaps and hence glycosylation sites were not predicted.

comparable number of glycosylation sites compared to most mammalian OVGP1, suggesting CHIO orthologues exhibit intermediate characteristics of CHIA (presence of catalytic motif) and OVGP1 (highly glycosylated tail).

6.8. Multiple sequence alignment of vertebrates GH18 proteins

In order to explore the similarities and differences in primary structure of the mammalian GH18 proteins, multiple amino acid sequence alignments (excluding CTBS and CHID1) were constructed by selecting one representative from each paralogue (Figure 6.10; Table 6.2). For convenience, herein the amino acids positions of compared proteins are indicated with reference of human CHIT1. The catalytic site residues and cleft lining aromatic residues are conserved to varying degree in the human and mouse GH18 proteins (Table 6.2). Among the important observations gathered: the catalytic active sites between residues 115-119 (DXDXE) is strongly conserved in both known active chitinases (CHIA and CHIT1) and the newly identified paralogue (CHIO). However, in CTBS, Asp115 is substituted with Asn102. With the exception of the complete substitution of Glu119 in all ChiLs, Asp115 and Asp117 are variably substituted among the compared homologues (Table 6.2). The ligand binding cavity of chitinases and ChiLs is lined with the aromatic residues (Fusetti *et al*., 2002; Houston *et al*., 2003; Olland *et al*., 2009; Schimpl *et al*., 2012). These residues are largely conserved among all the compared paralogues. For example Tyr/Phe191, Tyr/Phe246 and Trp337 were found strictly conserved across all chitinases and ChiLs. Trp10 is conserved in all homologues except Chil7 where Val substituted it at the corresponding position. Among paralogues, Trp197 is only substituted with different residues in three mouse ChiLs. Tyr169 has been reported as a residue lining the ligand binding cavity of CHIT1 (Fusetti *et al*., 2002), however in the sequence alignment it did not reveal any noticeable pattern of conservation. Trp78 is also conserved except in CHIL2 and Chil7 where it is replaced by Tyr and Arg respectively. Tyr13 and Trp50 were also appeared conserved in all active chitinases, CHIL1 and BP40.

In mammalian CHIA, a strictly conserved triad of three residues; Arg124, His187 and His248 have been attributed to its optimal activity in acidic pH (Olland *et al*., 2009). The triad is completely conserved between mouse and human orthologues of CHIA (Table 6.2). In addition, Arg124 is conserved in most homologues except CHIT1, CHIL2, Chil8, CTBS and CHID1. His 187 is completely unique to CHIA whereas His248 is only conserved in CHIA, CHIL2 and murine specific Chils except Chil5. Previously, it has been proposed that CHIL1 may bind with the heparin (Fusetti *et al.*, 2003) owing to the presence of a positively charged residue cluster (GRRDKQH) in the region: 122-128. In CHIL2 the corresponding region is less basic and contains only two positively charge residues, Lys and His at the corresponding positions. While in the other chitinases and chilectins this region lacks conservation. Since CHID1 and CTBS have evolved as separate GH18 gene lineages and share very low sequence identity with other GH18 homologues (17% and 9% respectively with human CHIT1), to attain maximum accuracy in the alignment of

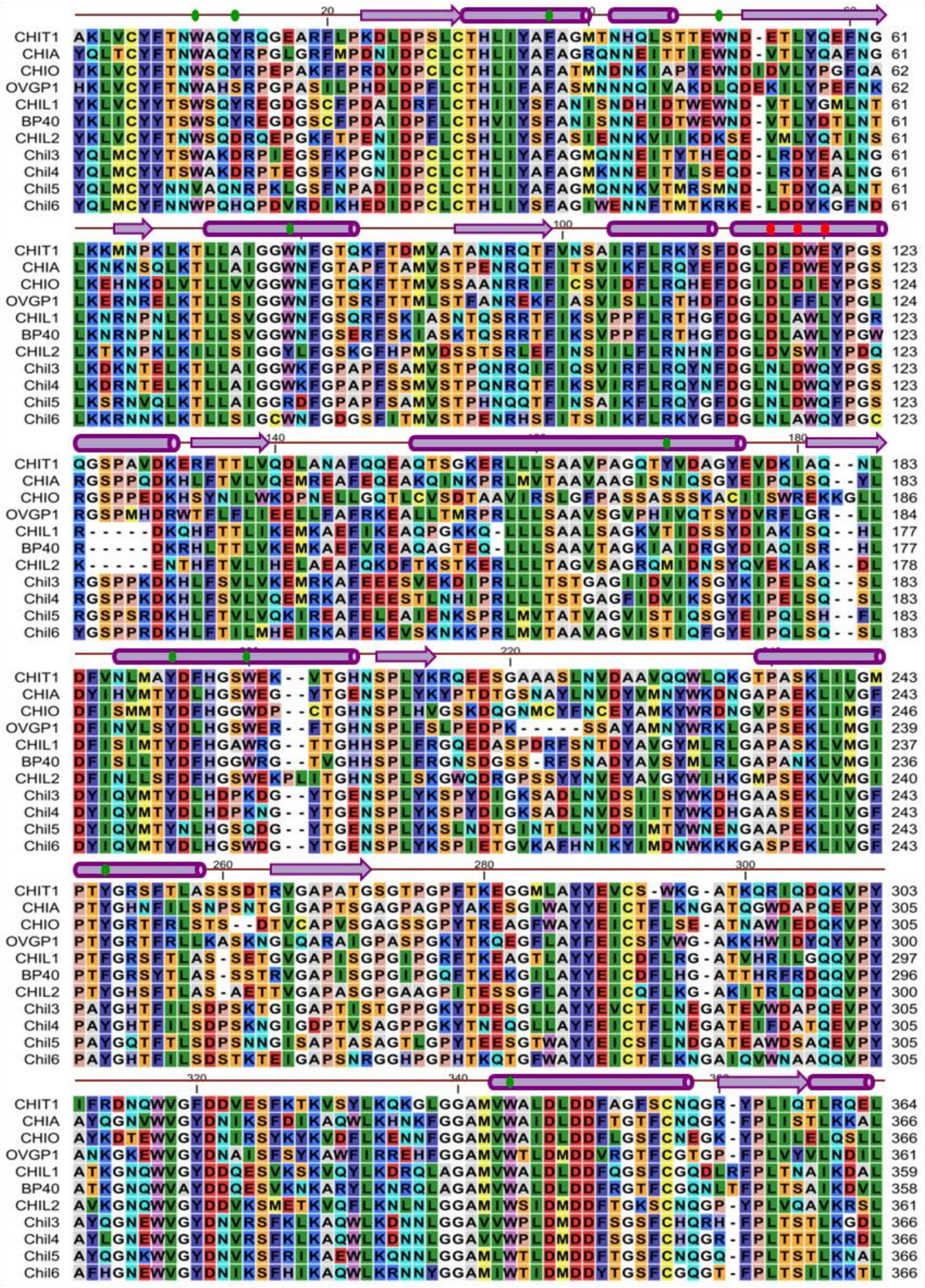


Figure 6.10. Multiple sequence alignment of GH18 proteins. A multiple sequence alignment of GH18 proteins (one per paralogue excluding CTBS and CHID1) is shown. The N-terminal leader sequence and C-terminal extension (where present) were removed. Secondary structure with reference to human CHIT1 structure (PDB id: 1LQ0) is schematized using purple cylinders (α helices) and arrows (β strands) at the top of each respective row. Catalytic site residues are indicated with red filled circles whereas aromatic residues aligning the ligand binding groove are indicated by green filled circles. All sequences are of human origin except Chio and BP40 from cow and Chil3, Chil4, Chil5 and Chil6 from mouse.

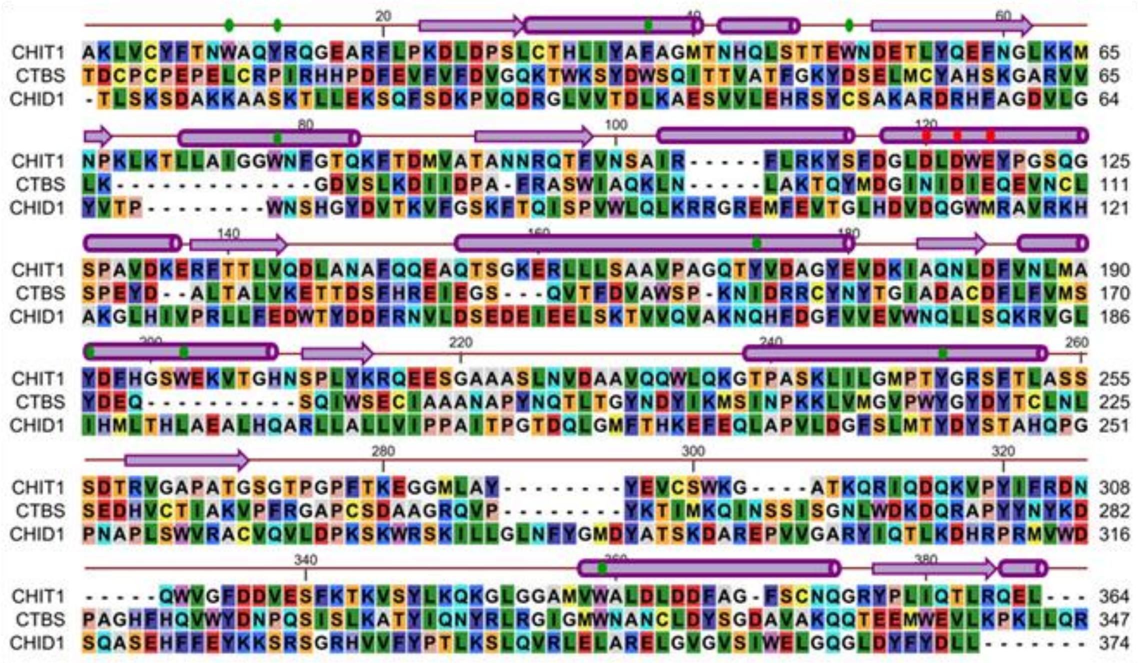


Figure 6.11. Multiple sequence alignment of CHIT1, CTBS and CHID1. A multiple sequence alignment of human CHIT1, CTBS and CHID1 is shown. The N terminal leader sequence was removed and the secondary structure with reference to the CHIT1 structure (PDB id: 1LQ0) is delineated using purple cylinders (α helices) and arrows (β strands). Catalytic site residues are represented with red filled circles and aromatic residues aligning the ligand binding groove are represented by green filled circles

Proteins	Catalytic Residues			Aromatic Residues (Lining the ligand binding groove)										Residue for Optimal acidic pI		
CHIT1	D115	D117	E119	W10	Y13	F37	W50	W78	Y169	Y191	W197	Y246	W337	Q124	N187	R248
Chit1*	D115	D117	E119	W10	Y13	F37	H50	W78	L169	Y191	L197	Y246	W335	R124	N187	R248
CHIA	D115	D117	E119	W10	Y13	F37	W50	W78	N169	Y191	W197	Y246	W338	R124	H187	H248
Chia*	D115	D117	E119	W10	Y13	F37	W50	W78	N169	Y191	W197	Y246	W339	R124	H187	H248
CHIO	D116	D118	E120	W10	Y13	F37	W50	W79	S170	Y194	W200	Y249	W339	R125	S190	R251
OVGP1	D116	F118	L120	W10	S13	F37	L50	W79	I170	Y192	W198	Y242	W334	R125	N188	R244
Ovgp1*	D116	F118	L120	W10	S13	F37	L50	W79	I170	Y192	W198	Y242	W334	R125	N188	R244
CHIL1	D115	A117	L119	W10	Y13	F37	W50	W78	T163	Y185	W191	F240	W331	R124	S181	R242
Chil1*	D116	A118	L120	W10	Y13	F37	W51	W79	A164	Y186	W192	F241	W332	R125	N182	K243
BP40	D115	A117	L119	W10	Y13	F37	W50	W78	A165	Y185	W191	F239	W330	R124	S181	R241
CHIL2	D115	S117	I119	W10	D13	F37	K50	Y78	M166	F186	W192	Y243	W334	K124	N182	H245
Chil3	N115	D117	Q119	W10	D13	F37	E50	W78	V169	Y191	K197	Y246	W339	R124	Q187	H248
Chil4	N115	D117	Q119	W10	D13	F37	E50	W78	V169	Y191	K197	Y246	W339	R124	Q187	H248
Chil5	N115	D117	Q119	V10	N13	F37	M50	R78	T169	Y191	Q197	Y246	W339	R124	Q187	Q248
Chil6	N115	A117	Q119	W10	H13	F37	R50	W78	T169	Y191	W197	Y246	W339	Y124	Q187	H248
CTBS	N102	D104	E106	--	--	F23	W37	--	N148	Y171	W178	Y216	W316	L111	F167	Y218
Ctbs*	N102	D104	E106	--	--	F23	W37	--	R148	Y171	W178	Y216	W316	S111	F167	Y218
CHID1	V171	E173	W175	W69	--	--	--	--	F222	Y242	W258	Y283	W361	--	S238	S283
Chid1*	V171	E173	W715	W69	--	--	--	--	F222	Y241	W258	Y283	W361	--	S238	D286

Table 6.2. Amino acids conservation of GH18 proteins. The table shows the conserved residues in different mammalian GH18 homologues on the basis of multiple sequence alignment and/or structural alignment. Residues at three functionally important regions namely, catalytic active sties, ligand binding groove aromatic residues and residues important for optimum acidic pI were compared. Conserved residues with reference to human CHIT1 and/or CHIA are shaded. Mouse orthologues are indicated by (*).

functionally important amino acids, CTBS and CHID1 proteins were examined independently using a structural alignment with human CHIT1 (Figure 6.11 and Table 6.2). This suggests the presence of catalytic activity in at least CTBS. Additionally, structural conservation of residues lining the ligand binding groove is mainly restricted to the last four residues (Y191, W197, Y246 and W337) which are present at the C-terminal region of both CTBS and CHID1.

The data not only show the absence of the catalytic motifs and the catalytic activity in the ChiLs but also provide information about the differences in the amino acid composition of the ligand binding cleft, suggesting differences in ligand specificity. Additionally, it is important to note that amino acids present at the proximity of cleft lining residue (except Trp337) vary considerably between the compared sequences. Such differences may affect the orientation of side chains of important aromatic residues and consequently alter the size and shape of the cavity.

6.9. Structure of chitinases and chitinase like proteins

In order to examine the effect of sequential conservation and variations on the three dimensional conformation of the human and mouse GH18 proteins, the structurally unresolved human and mouse chitinases and ChiLs, were modelled. Models of human OVGP1 and CTBS, mouse Chit1, Chia, Ovgp1, Chil1, Chil4, Chil7, Chil8, Ctbs and Chid1 and cow Chio (the latter not present in human and mouse) were developed using the closest available templates (for mouse Chils; PDBid 1VF8 , for human OVGP1; PDBid 3FXY, human CTBS; PDBid 3XY, 1LQ0 , cow Chio; PDBid 3FXY) based on the maximum primary sequence identity or multiple threading in the case of CTBS and full length CHIT1 and CHIA. The final models were selected on the basis of minimized DOPE and structural constraints. Ramachandran plot analyses of the selected models show that more than 98% of the residues of the modelled structures were present in the allowed region, supporting the structural plausibility of the molecular models (Figure 6.12). Moreover, Q mean scores of all selected models range between the acceptable limits of 0.0-1.0. The vertebrate GH18 proteins are typically composed of two structural domains: a core domain which adopts a TIM barrel $(\beta/\alpha)_8$ conformation and a relatively small second domain, situated between β7 and α7 called the α+β domain (Sun *et al*., 2001; Fusetti *et al*., 2002; Olland *et al*., 2009; Houston *et al*., 2003; Meng *et al*., 2010). Evaluation of all the modelled structures revealed that both domains are present in nearly identical spatial positions in relation to the known structures (Figure 6.13). In order to further examine the structural similarities, all the modelled and structured mammalian chitinases and ChiLs were superimposed in two sets: one set comprising all the structures with the exception of CTBS and CHID1 and the second set include CHIT1, CTBS and CHID1.

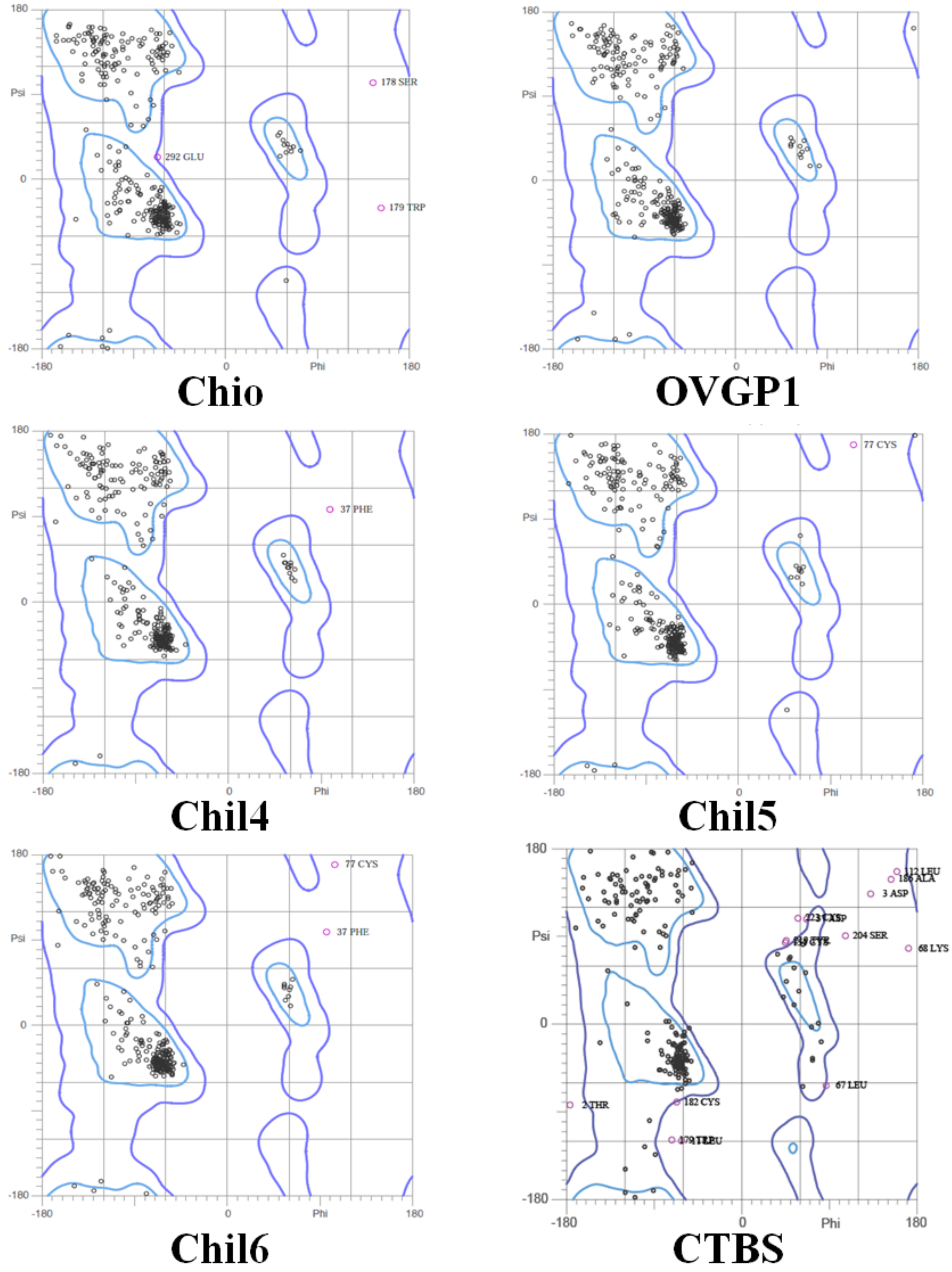


Figure 6.12. Ramachandran plots of modelled GH18 proteins. Ramachandran plot analyses of the modelled proteins were conducted using Molprobity server. The X and Y axis of each plot represent phi and psi angle respectively. Each black dot is the representation of the position of an amino acid with reference to the ratio of phi and psi angles. Light blue contour margins indicate the strictly allowed region while dark blue contour lines represent the generously allowed regions in the plot. Note all models are from the mouse sequences except Chio (from cow) and OVGP1 (from humans).

Additionally, mouse Chit1, Chia, Ovp1, Chil1, Ctbs and Chid1 were individually superimposed over the respective human orthologues (Figure 6.14). In set 1, superimposition of molecules showed that there is considerable conservation, not only in the Cα backbone architecture (RMSD values <0.5), but also in the distribution and spatial positioning of the secondary structural elements. The only exception in this regard was the modelled structure of cow Chio, where two small loops (Lys173-Ile176 and Thr257-Pro264) did not coincide with the loops of other molecules at the corresponding position (Figure 6.13 A&B). Conserved cysteine residues which form disulphide bonds in CHIT1, Cys5-Cys30 and Cys286-Cys349 were observed in all the structured and modelled proteins. Two overlapping antigenic epitopes described for CHIL1 (Pro238-Glu250 and Arg242-Gly252) (Boots *et al*., 2007) occupy the same spatial positions and structural conformation (β_8) in all the compared proteins. By comparison, CHIT1, CTBS and CHID1 tertiary structure superimposition revealed variation in the spatial arrangement of both the Cα backbone and secondary structural elements (Figure 6.13 C&D). Nevertheless, the overall topologies of the proteins are highly similar in terms of arrangement and positioning of domains (TIM barrel and α+β). Superposition of the mouse models over the respective human orthologues did not reveal any substantial differences in the conformation and spatial positioning of the secondary structural elements indicating high structural conservation between orthologues (Figure 6.14). For the mammalian chitinase and CHILs proteins where resolved structures are available, only the 39kDa GH18 region (including the TIM barrel and α+β domain) has been resolved, while the C-terminal tails of CHIT, CHIA and full length OVGP1 have not been structured. In order to explore the contribution of this domain to the structure, full length models of these proteins were generated (Figure 6.15). The models reveal that these C terminal extensions (CBM14 in case of CHIT1 and CHIA; Mucin like tail in OVGP1) form spatially discreet structurally unfolded loops. In the case of CHIT1 and CHIA the tails runs at 90^o to the active site groove “Scorpion like” and it can be imagined how this might contribute to the binding of a chitin chain, perhaps even guiding it into the active site groove, In the case of OVGP1, the tail extends from the rear (fish like) of the protein (in relation to the active site groove).

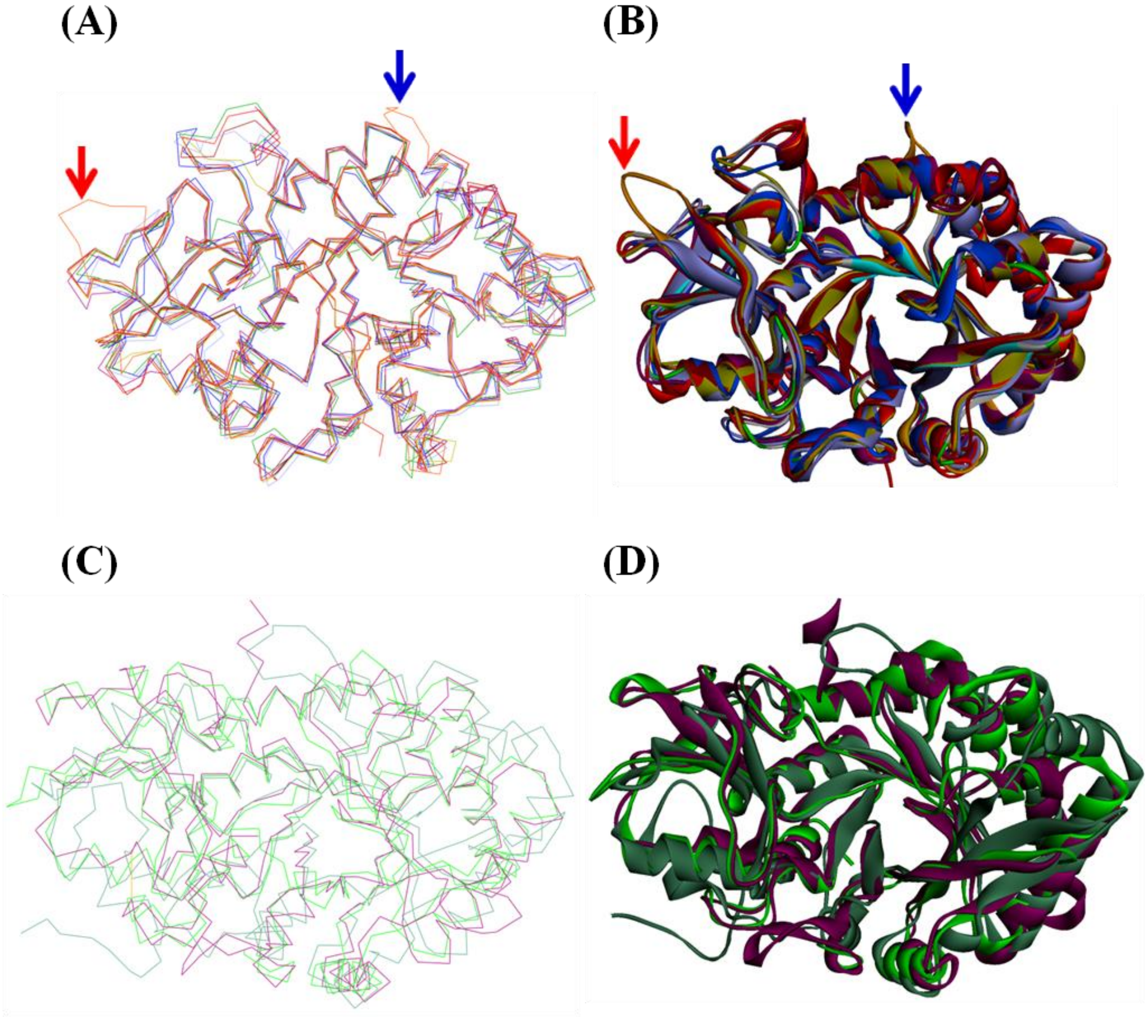


Figure 6.13. Superimposition of GH18 Proteins Structures. The resolved and modelled proteins have been superimposed over each other in two sets. The first set includes a human, mouse (except Chil1 and nearly identical murine Chil4, 5 and 6) and cow representatives of each of the mammalian GH18 paralogous proteins, excluding CTBS and CHID1 (A&B). The second set shows the superimposition of human CHIT1, CTBS and CHID1 (C&D). Superimposition of the Cα back bone is shown in **A** and **C**, while **B** and **D** show the secondary structure superimposition. Each molecule is coloured differently: human CHIT1 (purple), human CHIA (red), cow CHIO (orange), human CHIL1 (blue), cow BP40 (light blue), human CHIL2 (green), human OVGP1 (yellow), mouse Chil3 (brown), human CTBS (light green) and human CHID1 (sea green). Note variation in the Cα backbone and secondary structure in **C** and **D** (CHIT1, CTBS and CHID1) in comparison to A and B (all GH18 homologues excluding CTBS and CHID1). The differently placed cow Chio is shown in blue (Lys173-Ile176) and red (Thr257-Pro264) arrows.

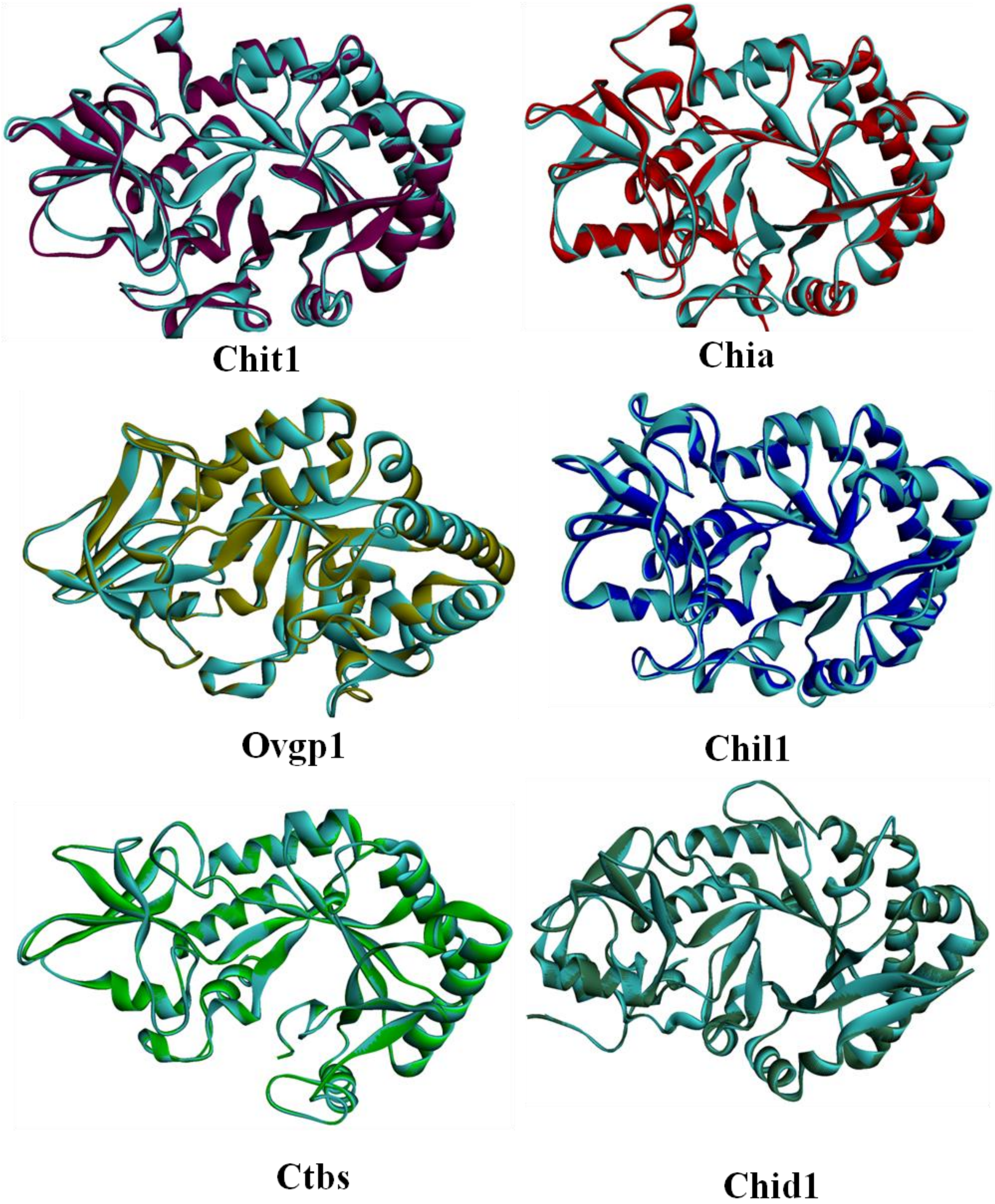


Figure 6.14. Superposition of human and mouse orthologues. Ribbon diagrams of human and mouse GH18 homologues were superimposed according to their orthologous relationship. Human models are differently coloured while all mouse homologues are coloured cyan.

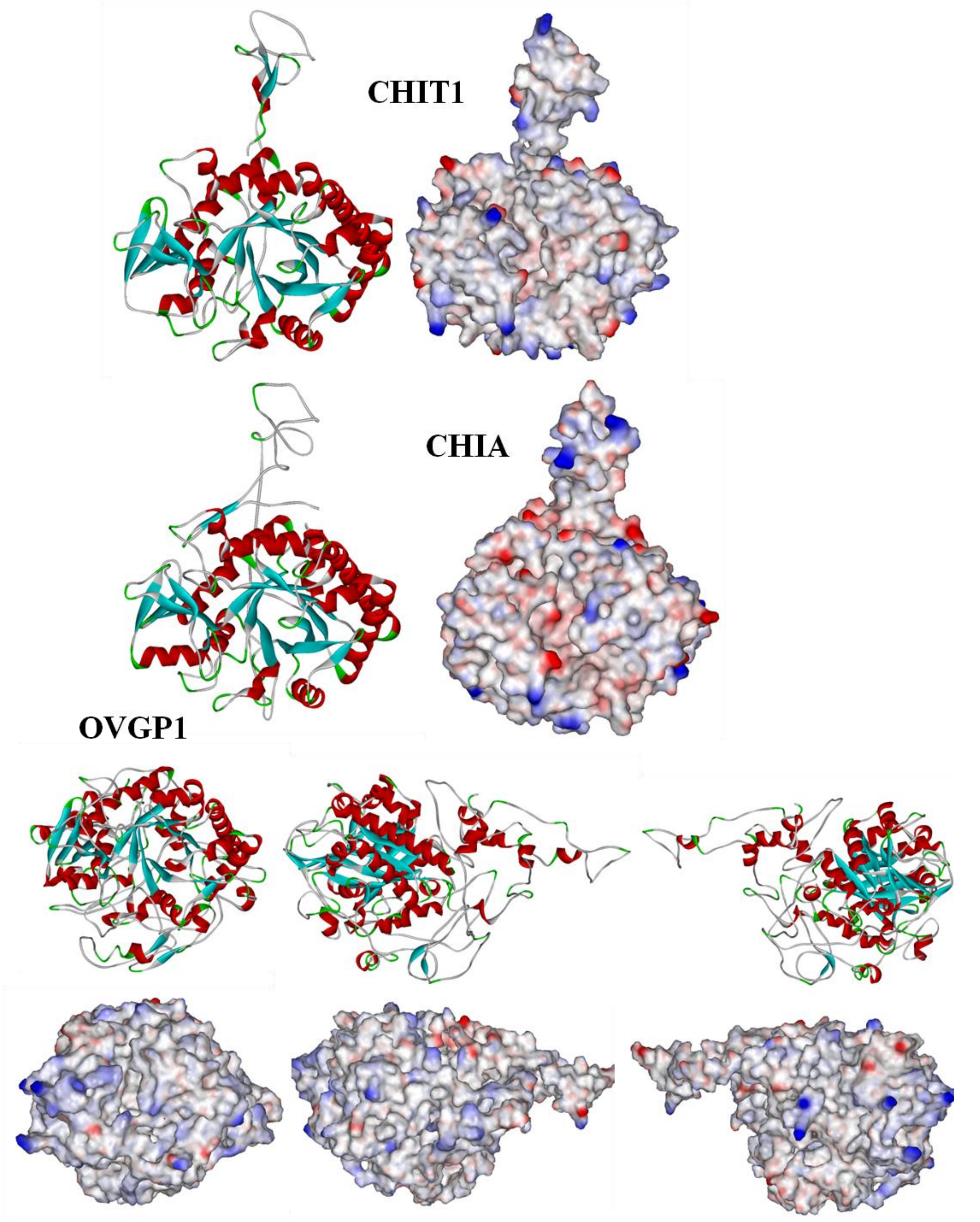


Figure 6.15. Full length models of human CHIT1, CHIA and OVGP1. Ribbon diagrams and electrostatic surface models of full length human CHIT1, CHIA and OVGP1 are shown. Full length models were developed by I-TASSER, where the GH18 domain was constructed using the corresponding templates while the C-terminal region is mostly developed by *ab initio* modelling.

6.10. Variations in ligand binding grooves of different GH18 paralogues.

Although strong structural similarities reflect the shared ancestry of the mammalian GH18 proteins, it provides limited insights for understanding the ligand specificity of these paralogues. To explore this further, the biophysical and structural characteristics of the ligand binding cavities were compared. To identify the potential cavities, the protein examined using POCASA 1.0. The program scans the protein surface in a 3D grid and by placing spherical probes (radius 2.0Å) in cavities it delineates the dimensions and depth of potential ligand binding cavities and ranks them according to volume (i.e. the number of accumulated probes (Table 6.3; Figure 6.16). Consistent with earlier studies (Fusetti *et al*., 2002; Houston *et al*., 2003; Olland *et al*., 2009; Meng *et al*., 2010) the central cleft was detected in all the modelled structures, pointing to a similar ligand binding region. However, estimation of cavity volume revealed considerable variation in the size, shape and depth of these grooves. Among the expected GH18 the largest ligand binding groove was observed in human CHIT1 followed by human CHIL1 and mouse Chil3 and 4. Intriguingly, despite the close evolutionary relationship of BP40 with CHIL1 the cleft (groove) volume of BP40 is notably (1.5 times) smaller (358Å^3) than human and mouse CHIL1 (543Å^3 and 575Å^3 respectively). Conversely, despite the distant evolutionary relationship of CHIL2 with CHIA their ligand binding clefts are of similar volume (431Å^3 and 426Å^3 respectively). Active groove of mice shows a smaller volume then human CHIT1, but is similar in size to the groove volume of Chia of both species (Table 6.3; Figure 6.16). Mouse and human CHID1 reveal the smallest groove of the molecules examined, with other cavities on the surface having greater volume. The groove cavities of CTBS is amongst the largest (1164Å) observed in human GH18 proteins. The cavity opens out to a greater degree than the other molecule at lower end of the groove (as visualized in figure 6.16).

The shape of the ligand binding region generally defines the nature and type of ligand that binds to the protein. Despite the similarity in the volumes of the grooves between CHIT1, CHIL1 and Chil3, the shape of their cavities differs considerably. The central groove of CHIT1 and CHIL1 runs down the length of the molecule giving it a tunnel like appearance. Similarly, the shape of CHIA, CHIL2 and Chio also appears to be elongated and tunnel like in appearance suggesting that these proteins can bind with an oligomeric (carbohydrate) moiety. The ligand binding cleft of Chil6 appears irregular in shape. By comparison, the Chil5 central cleft shape shows more resemblance to the antecedent homologue CHIA (Fig 6.16). This suggests that the different rodentia specific chilectins may have different ligand specificity. Interestingly, the central ligand binding groove of the active chitinase CTBS is different from CHIT1, CHIA and CHIO, with a wide and

Protein	PDB ID	Volume (Å^3) Human	Volume (Å^3) Mouse
CHIT1	1LQ0	644	472
CHIA	3FXY	426	474
Chio*	(3FXY)	364	---
OVGP1*	(3FXY)	378	343
CHIL1	1NWU	543	575
BP40	2ESC	135	---
CHIL2	4AY1	431	---
Chil3	1VF8	---	551
Chil4	(1VF8)	---	473
Chil5*	(1VF8)	---	403
Chil6*	(1VF8)	---	389
CTBS*	(3FXY, 1LQ0)	1164	500
CHID1	3BXW	34	29

Table 6.3. Volume of Ligand Binding Grooves. PDB coordinates of all structured and modelled (*) proteins were submitted to POCASA 1.0 in order to estimate the volume of the central ligand binding groove with 2Å (radius) spherical probes. PDB ids of the structurally known GH18 molecules or templates used to model the unresolved proteins are given.

irregular shape which possibly reflects its exo chitinase catalytic activity. The central cleft cavities of human and mouse orthologues are similar to each other. As mentioned earlier CHIL1 and BP40 evolved by the duplication of a common ancestor but the ligand binding groove of BP40 is smaller than CHIL1 and ranked as third largest as compared to other cavities present in the same protein. The first (177Å^3) and second (157Å^3) ranked cavities are present near the start (α_9) and end of the α+β domain (between β_{10} and β_{11}) respectively. Similarly, the central groove of CHID1 was also ranked as smallest among the potential cavities. An electrostatic surface analysis revealed that interiorly, the central cavity is hydrophobic in nature in all GH18 homologues except in CTBS where two of polar residues (Arg and Asn) were found buried inside the ligand binding groove (Figure 6.17). Additionally half of the rim around the central binding groove is relatively enriched with polar residues, which may define the orientation of the ligand. Comparison between the mouse homologues with the comparable human GH18 proteins did not show any significant differences in the shape and electrostatic surface (Figure 6.18). This points the similarities in ligand specificity and potentially functions between the compared orthologues.

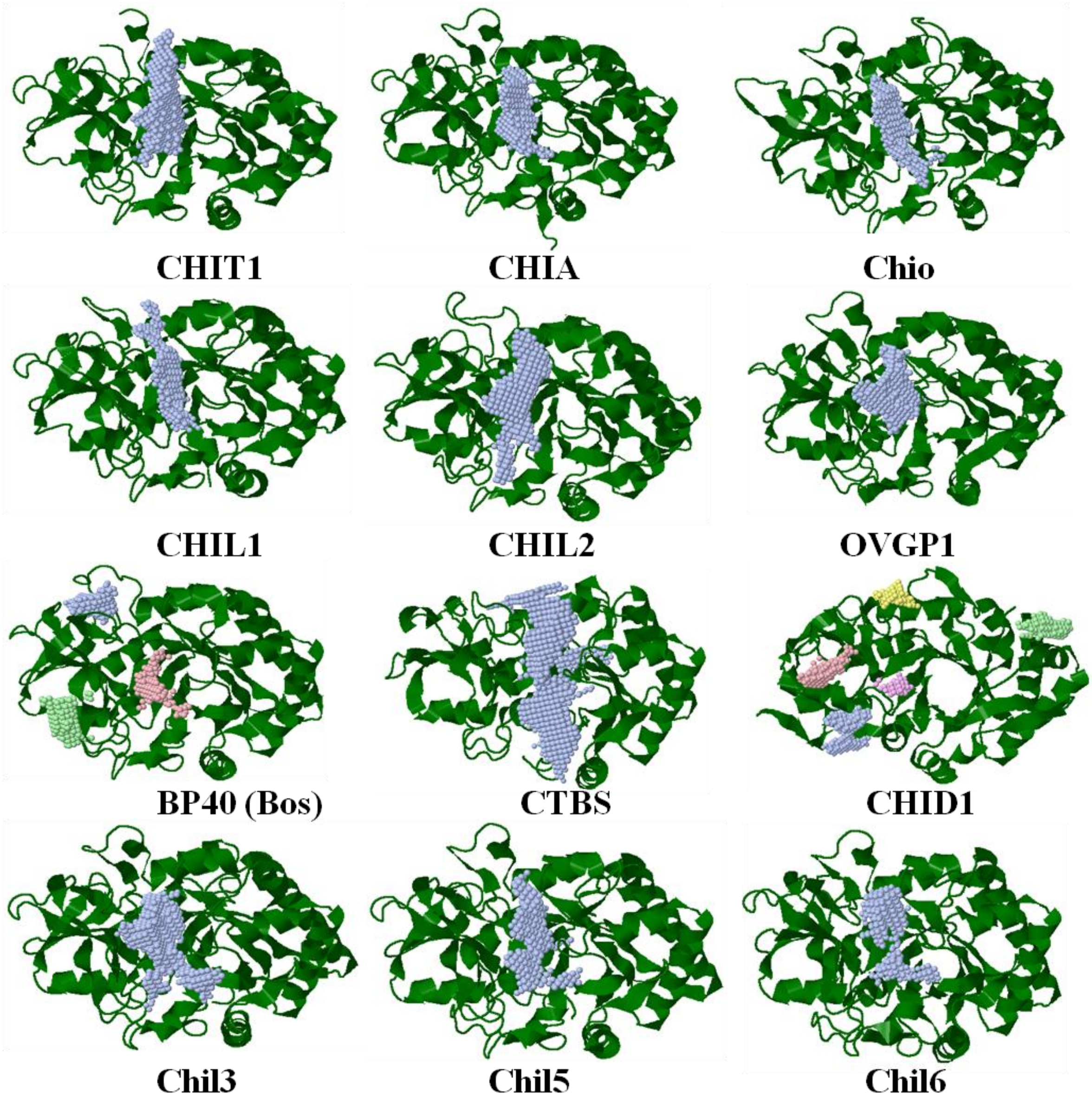


Figure 6.16. Volume of Ligand Binding Groove. PDB coordinates of all structured and modelled GH18 homologues were submitted to POCASA 1.0 to evaluate the volume of potential ligand binding grooves with 1Å spherical probes. The protein molecules are represented with green ribbon while the density of probes are coloured as light blue, light green, pink and yellow in decreasing order of volume of cavities.

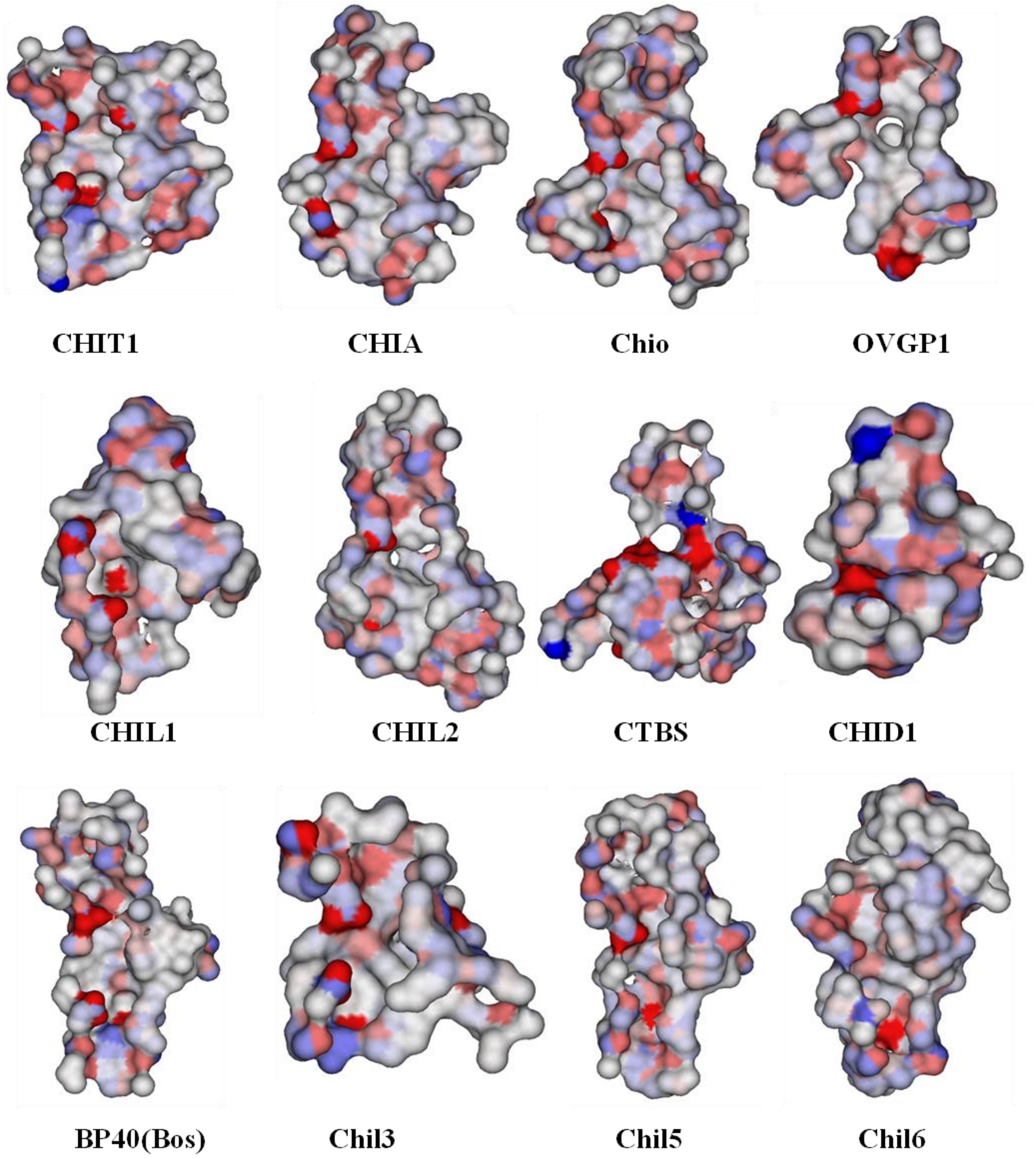


Figure 6.17. Electrostatic surface of the ligand binding groove. The electrostatic surface of the ligand binding groove as visualized by DS visualizer 3.5 is shown here. Red and blue refers to the negatively and positively charged residues respectively. Note the irregular margins of the cleft in Chil3, Chil6, OVGP1, CTBS and CHID1. Also to note, the division (compartmentalization) of the central groove into smaller pockets in the case of BP40 and Chil6.

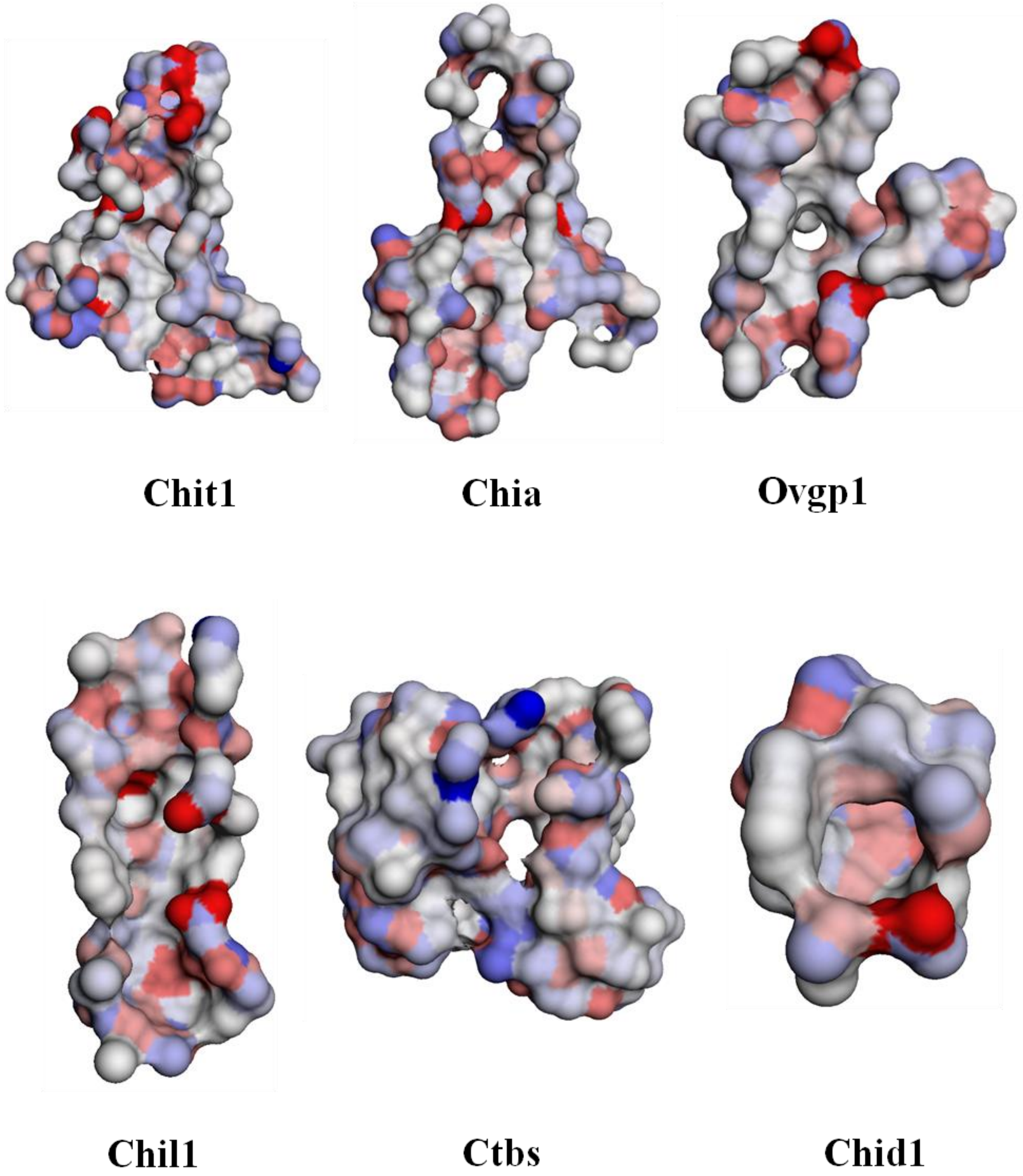


Figure 6.18. Electrostatic surface of ligand binding cavities of mouse GH18 proteins. Electrostatic surface of ligand binding groove as visualized by DS visualizer 3.5 is shown here. Red and blue refers to the negatively and positively charged residues respectively.

The ligand binding grooves of the resolved GH18 homologues are lined with solvent exposed aromatic residues (Sun *et al*., 2001; Fusetti *et al*., 2002; Houston *et al*., 2003; Olland *et al*., 2009). Superimposition of these aromatic amino acids revealed significant similarities in the spatial orientation of these residues between different proteins (Figure 6.19). However, some notable differences were observed, for instance Trp78 of CHIL1 and BP40 and Trp197 of CHIL1 showed a different orientation in comparison to the other proteins. Although Trp337 is completely conserved in all the compared proteins, it is slightly tilted in case of Chil3 and Chil4 molecules. Similar comparison of CHIT1 with CTBS and CHID (Figure 6.20) did not demonstrate such similarities in the orientation of the lining residues, with the exception of Tyr191, Trp197, Tyr246 and Trp339.

Taken together the variations in shape and size of the ligand binding cleft, and the polarity and orientation of aromatic residues indicates that there is likely to be differences in the ligand binding specificity among different mammalian GH18 paralogues. Such differences in the active sites will reflect new and/or additional functional role(s) among vertebrate chitinases and ChiLs since their evolution from the ancestral gene.

6.11. Summary of findings

- The human genome encodes seven members of the GH18 protein family, these include three active chitinases (CHIT1, CHIA and CTBS) and four chitinase like proteins or chilectins (CHIL1, CHIL2, OVGP1 and CHID1).
- Phylogenetic analyses demonstrated three major evolutionary lineages of GH18 proteins which diverged prior to the origin of vertebrates, leading to the formation of CHID1, CTBS and other GH18 homologues.
- At the base of the vertebrate tree, extensive expansion of the GH18 genes (excluding CTBS and CHID1) were observed leading to the formation of ancestral CHIT1, CHIA and CHIO (a newly identified paralogue in this study).
- The CHIT1 ancestral gene duplicated before the reptilian-mammalian split and again at the root of mammals leading to formation of two chilectins CHIL2 and CHIL1 respectively.
- The CHIA ancestral gene duplicated extensively in vertebrates leading to the formation of two active chitinases at the root of vertebrates and one at the base of primate divergence.
- With the origin of mammals the CHIA ancestral gene duplicated to form a chilectin, OVGP1, and in family *muridae* additional array of chilectins.

- Synteny analyses of human and other vertebrate genomes recapitulate the phylogenetic association observed for the genes, where CHIA and CHIT1 related genes are closely linked. The only exception in this regard is the proximity of chil2 to chia, reflecting a genomic arrangement prior to the physical separation of the chia and chit1 ancestors in mammals.
- Several GH18 protein encoding genes were identified in the fishes. With the exception of a few CHIA orthologues, the orthologous relationship of most homologues is ambiguous. A separate phylogenomic analysis suggested that these are potentially members of CHIT1 and CHIO groups. Additionally, relics of the fish specific genome duplication were also observed.
- Potential relics of gene “death” as pseudogenes of CHIA and CHIO were found in the human genome.
- Structurally, all GH18 homologues (human and mouse) are strongly conserved, comprising two distinct protein domains, a TIM barrel and $\alpha+\beta$ domain.
- The CBM14 tail of CHIT1 and CHIA and highly glycosylated tail of OVGP1 are predicted to be unstructured.
- Differences in the volume and shape of the central cleft were observed in different GH18 paralogues, which may determine the size and nature of the ligand for the proteins. The data suggest that in contrast to human chilectins, mouse Chil3 and Chil4 and bovine BP40, may not be able to bind with oligomeric carbohydrate moieties.

6.12. Discussion

Multiple gene duplications and a differential rate of gene death across species lineages provide obstacles in inferring the evolutionary history of gene families (Roy, 2009). Moreover, inaccurate gene annotation may further confound the issue (Bidartondo, 2008; Demuth and Hahn, 2009). Hence a combined approach of data mining, phylogenetics and comparison of genomic synteny can provide a better resolution to not only annotate the gene sequences but also to identify the relationship of gene families with reasonable accuracy. In the present study we have exploited all three approaches to understand the phylogenetic relationship of GH18 family members in the vertebrate lineage. Our analyses presents a more detailed view of the evolutionary dynamics of chitinase and ChiLs in vertebrate. In addition more paralogues within the family were identified. The data provide an improved time line approximation for the expansion of GH18 family genes in vertebrates. In an investigation to explore the variability in the protein structure and ligand

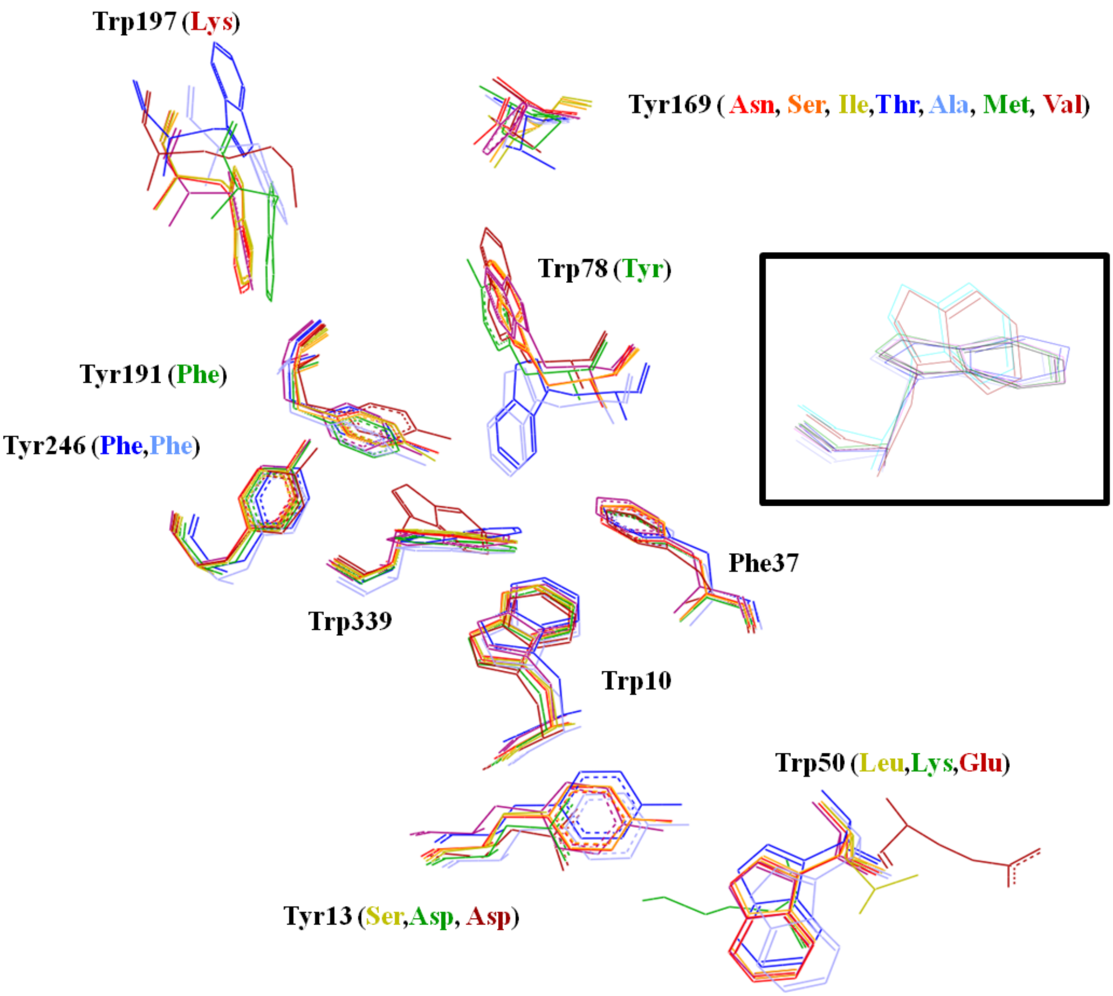


Figure 6.19. Superimposition of the aromatic residues lining ligand binding grooves. Aromatic residues of GH18 proteins (excluding Chil4, CTBS and CHID1) lining the ligand binding groove were superimposed. In the inset magnified view of Trp337 superimposition is shown with all murine chilectins (Chil3, Chil4, Chil5 and Chil6) with human CHIL1. Note the tilt in the side chain of Trp337 in Chil3 and Chil4 in comparison to the other chitinases and ChiLs. Key: CHIT1 (purple), CHIA (red), OVGP1 (light green), CHIL1 (blue), CHIL2 (green) BP40 (light blue) and Chil3 (maroon).

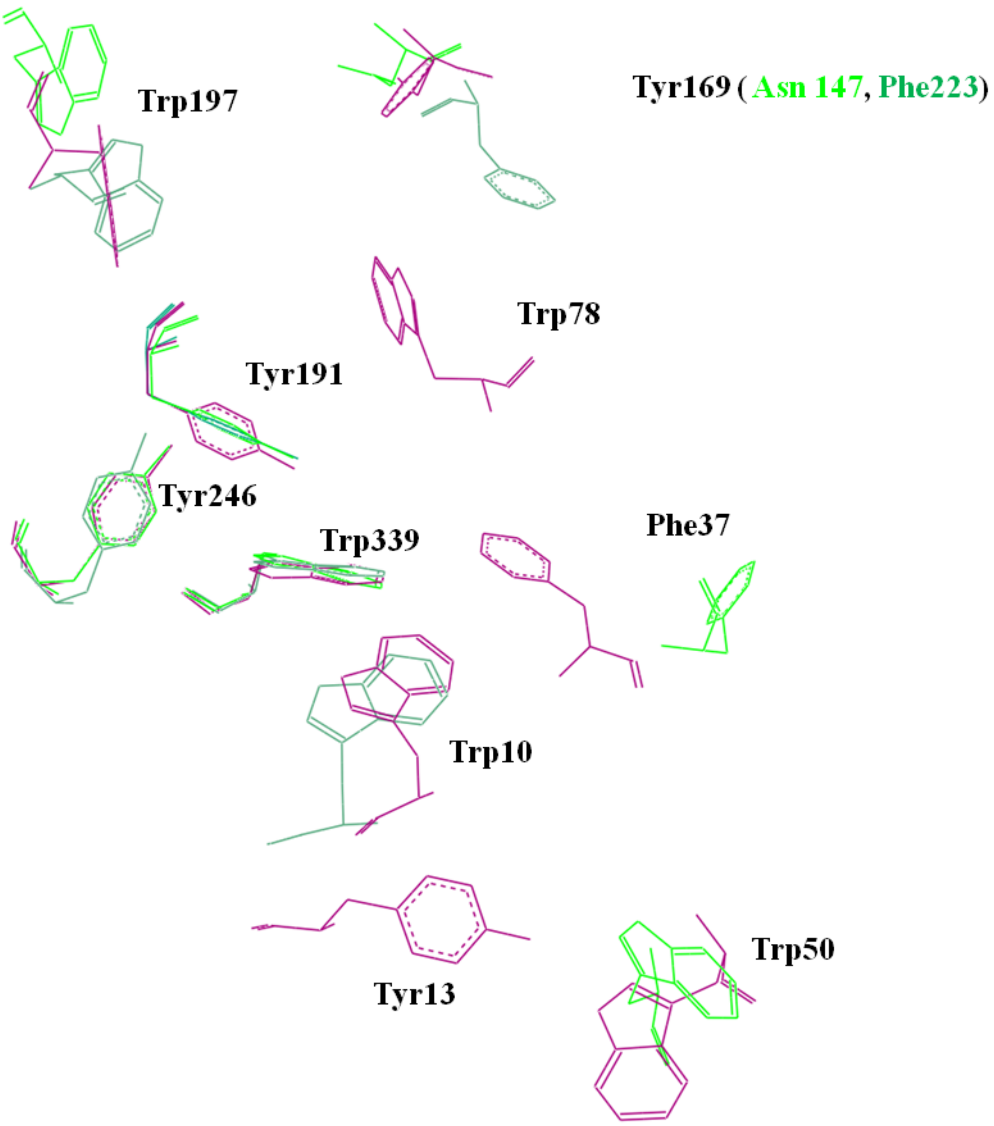


Figure 6.20. Superimposition of the aromatic residues lining ligand binding groove. CHIT (purple), CTBS (green) and CHID1 (cyan) aromatic residues lining the ligand binding groove were superimposed. Residues are numbered according to the CHIT1.

specificity, protein models and resolved structures of mammalian human, mouse and cow GH18 proteins were compared.

Evolution of GH18 proteins in vertebrates

The inferred evolutionary history is based on all the GH18 homologues of vertebrates and chordates revealed an early separation of CHID1 and CTBS from other GH18 genes, including homologues of lamprey, lancelet and tunicate, into separate clades. The tree indicated that CHID1 and CTBS shared a common ancestor to other GH18 genes, before the origin of chordates. Both CTBS and CHID1 genes were detected in almost all the vertebrates analysed, however with little evidence of gene death or gene birth in extant vertebrates. This sustained presence suggests that their functional role is vital for the organism's physiology. In contrast to CTBS and CHID1, other GH18 family members have undergone extensive gene expansion in the vertebrate lineage.

The present study is in agreement with the earlier studies (Bussink *et al.*, 2007; Funkhouser and Aronson, 2007) in demonstrating a duplication event that lead to the formation of two active chitinases, ancestral CHIA and CHIT1. Both genes subsequently underwent multiple episodes of gene duplication that lead to the formation of additional chitinases and ChiL genes. Previously it has been suggested that this initial duplication occurred in early tetrapods with the evolution of CHIA paralleling the development of an acidic stomach (Bussink *et al.*, 2007). However this study points to an event that occurred considerably earlier, before the divergence of bony fishes. Furthermore, this ancestral CHIA gene underwent multiple gene duplication events at different time points, in the evolutionary scale, of these at least two duplication events predate the divergence of bony fishes.

It is now widely established that two rounds of whole genome duplications referred to as 1R and 2R occurred before the divergence of chordates and vertebrates respectively. After these duplications, extensive gene loss and genomic rearrangements had occurred (Holland *et al.*, 1994; Putnam *et al.*, 2008; Hufton *et al.*, 2008). Genome maps of the last common chordate ancestor have been reconstructed and signature regions of 1R and 2R have been determined in different vertebrate chromosomes including human (Nakatani *et al*, 2007). Considering that the time of the gene duplication event that lead to the formation of the CHIT1 and CHIA ancestor predates the divergence of bony fishes, it is possible that the birth of CHIT1 and CHIA results from 1R or 2R. However, a comparison of the human CHIT1 and CHIA loci (chromosome 1 p and q arm respectively) with the proposed regions of 1R or 2R signature indicates otherwise suggesting this duplication event was independent of 1R and 2R. Therefore the simplest explanation is that the duplication leading to the formation of the CHIT1 and CHIA ancestors had occurred after 2R and before the divergence of bony fishes. Homologous sequences from the cartilaginous fishes (sharks and rays) once available may increase the precision of these estimates.

Orthology and paralogy of fish GH18 homologues

In comparison to mammalian and fishes GH18 sequences, fewer homologues are available from intermediate vertebrates (amphibian, birds and reptiles). Either for this reason or due to concerted evolution of fish sequences it was not possible to clearly establish the orthology of most fish homologues. However, some homologues unambiguously align to the CHIA clades (both CHIA1 and CHIA3). It is possible that fish homologues that are present at the base of the CHIA and CHIT1 superclades are CHIT1. Similarly, the distinct fish clades which are variably positioned within or at the base of the CHIA superclade may be CHIO genes. It is conceivable that both potential fish CHIT1 and CHIO may have undergone concerted evolution followed by the heterologous recombination and unequal cross over (Dover, 1982; Pinhal *et al*., 2011). This can result in the loss of the phylogenetic signal, to give ambiguous results in the trees.

Two distinct clades of potential fish CHIO sequences point towards the telost specific genome duplication (3R) which occurred around 350 mya (Meyer *et al*., 2005). Gene synteny analysis also strengthens this idea as paralogues present in the two chio clades are located on different chromosomes and at least in one case (medaka) these two chromosome are proposed to result from 3R (Nakatani *et al*., 2007). It has been suggested that most duplicated genes which arose as a result of whole genome duplications (1R, 2R or 3R) died out because of the redundancy (Lynch and Conery, 2000). However, duplicated genes of CHIA/CHIO and to some extent CHIT1 resulting from 3R persist in many of the compared extant fishes. This expansion followed by retention of genes could be explained in terms of evolutionary pressure for example provided by the high chitin containing diet (molluscs and crustaceans) (Fines *et al*., 2009), fungal pathogen in an aqueous environment and for some developmental necessities (Wagner *et al*., 1993).

Evolution of CHIA in vertebrates

Multiple gene duplications occurred in ancestral CHIA giving rise to CHIO, CHIAII and CHIAIII as well as OVGP1 and an array of ChiLs in family *muridae*. Of these the origin of CHIAII probably occurred at the base of primate lineage, while CHIAIII ancestor arose before the divergence of the bony fishes. In the human genome relics of both paralogues are present as pseudogenes. Duplication of ancestral CHIA in the family *muridae* resulted in the emergence of Chils which further duplicated to produce an array of Chils in mouse and rat. The underlying reason of this expansion is difficult to assess however, given the immune related function of Chil3 (Sutherland *et al*., 2011) it is reasonable to speculate that a rapidly evolving immune system may provide the suitable selection pressure for the emergence and retention of murine specific chilectins.

One of the most interesting observations of the phylogenetic analysis is the presence of an additional clade at the base of CHIA group, populated with amphibian, reptilian and mammalian sequences. We hypothesize that the gene duplication in ancestral CHIA to give rise to CHIO paralogues occurred before the divergence the bony fishes. From the topology of the phylogenetic trees it appears that the duplication of ancestral CHIA which resulted in the emergence of the OVGP1 paralogue, occurred before the divergence of bony fishes. However, no clear orthologue of OVGP1 was detected in any non mammalian vertebrate. This raises two possibilities that either OVGP1 indeed arose before the divergence of bony fishes and that all non mammalian vertebrates lost the OVGP1 gene, or the difference in the nucleotide substitution rate of OVGP1, compounded with the data bias (fewer non mammalian sequences) make the emergence of OVGP1 appear earlier than in the case. Indeed, the branch length of the OVGP1 suggests a relatively fast evolutionary rate in comparison to CHIA, resulting in the more distinct phylogenetic signals. OVGP1 is involved in fertilization and early embryonic development (Buhi *et al*., 2002) and accelerated rates of evolution have been observed in other genes (ZP2, ZP3, ADAMs2 and ADAMs32) associated with reproductive physiology (Swanson *et al*., 2001), therefore a later origin of OVGP1 than the tree suggest is plausible. The CHIO clade shows common ancestry with either CHIA and OVGP1, depending on the out group used (lancelet and nematode worm). Therefore based on the current evidence we propose that an initial duplication in the ancestral CHIA gave rise to the ancestor of CHIO and OVGP1 which then duplicated to form each gene, followed by faster OVGP1 evolution in mammals to give it a distinct clade topology.

Evolution of CHIT1 in vertebrates

The CHIT1 superclade shows that two main duplication event gave rise to the orthologues of extant CHIT1, CHIL1 and CHIL2. It was earlier proposed that gene duplication of ancestral CHIT1 resulted in the origin of CHIT1 and precursor of both ChiLs (Bussink *et al*., 2007; Funkhouser and Aronson, 2007). In contrast to this, the presented phylogenomics analyses indicate the first gene duplication event gave rise to CHIL2 and later another duplication in the CHIT1 gave birth to CHIL1. Furthermore the presence of the CHIL2 orthologue in the anole lizard points that this duplication event occurred prior to the origin of mammals.

In mammals, CHIT1 and CHIA are either present on different chromosomes or different arms of same chromosome with CHIL2 present in the proximity of CHIA and CHIL1 neighbouring CHIT1. Intriguingly, this physical separation is not evident in case of anole lizard as the CHIT1 homologue (partial sequence) is present next to CHIL2 in the animal. This observation suggests that the rearrangement of the CHIT1 gene occurred between the split of reptiles and mammals, before the diversification of the latter. As CHIL2 genes were identified in few mammalian species in comparison to CHIL1, it is possible that the function of CHIL2 has been taken over by CHIL1 rendering CHIL2 redundant and as a result CHIL2 is gradually being lost. Alternatively, in rodentia, the additional array of ChiLs may have evolved to replace the function of CHIL2 (Figure 6.21).

The information gathered from all phylogenomic analyses collectively suggest that genes of family GH18 have taken three evolutionary lineages before the emergence of vertebrates namely CHID1, CTBS and ancestral endochitinase. In the vertebrate lineage the ancestral

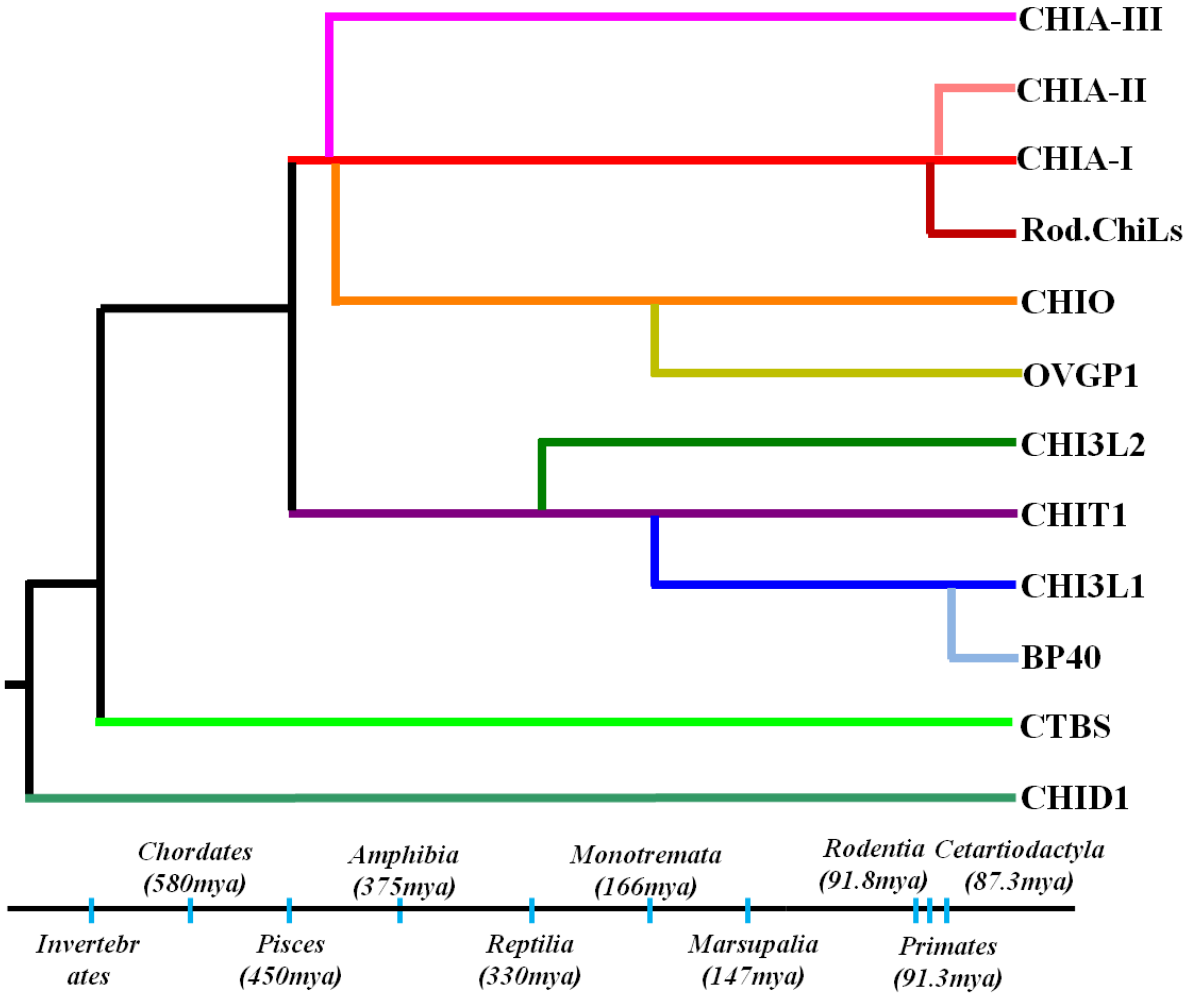


Figure 6.21. Composite model of evolution of vertebrate chitinases and chilectins. A schematic representation of gene duplication events (excluding most species specific duplications) leading to the mammalian Chi/ChiL genes is shown. The branch lengths are approximated to a time line (at bottom) showing million years ago (mya) (Bininda-Emonds *et al*., 2007). The loss of catalytic activity is shown by cross. Whole genome duplication events, 1R, 2R and 3R are indicated as red blue and green circles.

chitinase underwent extensive expansion which had lead to the emergence of CHIT1 and CHIA ancestors before the diversification of bony fishes around 450mya. Ancestral CHIA again underwent two events of gene duplication resulting in the birth of CHIA-III and CHIO/OVGP1 before the origin of tetrapods. Extensive gene death happened in both of these paralogues and relics of gene death are present in the form of pseudogenes in at least two primate species (human and macaque). Subsequently only one ChiL, CHIL2 arose in the common ancestor of reptiles and mammals (330mya) as a result of duplication of ancestral CHIT1. Two other ChiLs namely, OVGP1 and CHIL1 are the result of gene duplication of CHIA/CHIO and CHIT1 respectively which occurred before the divergence of mammalian lineages (166mya). In comparison to CHIL1 and OVGP1, CHIL2 underwent extensive gene loss in many mammalian lineages. Moreover, in some mammalian lineage GH18 genes underwent further gene expansion, such as in family *muridae* where duplication of CHIA resulted in the birth of an array of rodentia specific ChiLs and in family *bovidae* where duplication of CHIL1 gave rise to BP40.

Comparing the evolutionary history of genes to the structure function relationship of the encoded proteins provides insights into the underlying reasons in the expansion, diversification and retention of genes. For this purpose, the protein structure of selected examples was compared.

Structural conservation and potential ligands of GH18 paralogues

All the structured and modelled chitinases and ChiLs proteins share considerable similarity in their scaffolding. Without exception all of them exhibit distinct TIM barrel and α+β domains. This strong conservation in the overall structure and domain distribution reflects their common evolutionary origin. However, subtle differences in the potential ligand binding sites suggest functional innovations (neo/sub functionalization) during the evolution of GH18 proteins. Despite the strong tertiary structure conservation substitutions in the catalytic sites, variations in the amino acid composition, size and shape of the ligand binding groove indicate that not all chilectins could bind with oligomeric carbohydrate moieties of the same length. For instance the relatively irregular groove, combined with the loss of critical aromatic residues in mouse specific Chil3, suggests different ligand specificity to Chil1. Indeed, consistent with the unpublished data from our lab that showed that mouse Chil1 binds tightly with chitin beads while Chil3 and Chil4 showed relatively weak binding and readily dissociated from chitin beads in washing. This observation is in agreement with the earlier studies by Houston *et al*., 2003 and Tsai *et al*., 2004 which demonstrated poor binding of Chil3 with oligomeric carbohydrate ligands. In addition to the difference in the shape of the cavity it is possible that difference in the orientation of Trp339 between Chil1 and Chil3 may be a contributing factor in this regard because of its profound importance in ligand binding (Fusseti *et al*., 2002). A similar explanation could be extended to other closely related rodentia specific chilectins, Chil4 and Chil6 owing to the high amino acid identity (95% and 85% respectively) and structural similarity with Chil3. However, Chil3 and Chil4 differ in their expression pattern, as Chil3 has been found highly expressed in erythroblasts and bone marrow, whereas Chil4 expression has been noted in parotid gland and chondrocytes. This raises a possibility that these mouse specific ChiLs shares functional similarity but perform their role in different tissues or cell types. In contrast to other rodentia specific chilectins, the amino acid identity and ligand binding groove shape of Chil7 is more similar to Chia, suggesting that it may interact with a polymeric carbohydrate moiety. Although, CHIL1 and CHIL2 share a common evolutionary origin, out of 10 aromatic residues, 4 residues (lining ligand binding cleft) are substituted in CHIL2 and 2 of these substitutions are non iso-functional. Moreover, the expression patterns of CHIL1 and CHIL2 show different tissue specificity as in humans CHIL1 is predominantly expressed bone marrow, uterus and retina, while elevated expression of CHIL2 has been found in thymus and adipocytes (Hruz *et al.*, 2008). Considering, that structural and nutritional consumption of chitin is very limited in mammals and especially not known in humans, it has been proposed that human GH18 proteins may interact with the other carbohydrate moieties. Among these heparin sulphate, heparan sulphate and hyaluronic acid are frequently considered as most likely candidates. These sugars may serve as linking bridges between proteins, especially for the proteins which are involved in growth and proliferation (Plazinski and Knys-Dzieciuch, 2012; Schlessinger *et al.*, 2000). Although, to date, no direct empirical evidence has been demonstrated in this regard, functional studies of ChiLs suggest that they may interact with the several glycoproteins via their carbohydrate moieties. One such possibility is the binding of ChiLs with syndecan-1, a major cell surface proteoglycan, via its heparan sulphate linkage. This binding along with integrin αvβ3 leads to the phosphorylation of focal adhesion kinase (FAK). Phosphorylated FAK in turn leads to the activation (phosphorylation) of mitogen activated protein (MAP) kinase (Erk), phosphotidyl inositol 3 kinase (PI3K) and Akt Kinase. These molecular events instigate cell migration (transcytosis), proliferation, survival and angiogenesis, producing a supportive environment for oncogenesis (Alexopoulou *et al.*, 2007; Shao *et al.* 2009). Indeed elevated expression of ChiLs has been reported in several cancers in humans (Coffman, 2008; Zhu *et al.*, 2012a&b) and mouse models (Qureshi *et al.*, 2011). Studies have been shown that syndecan-1 and MAPK mediated molecular events are involved in the tissue modelling (Peretti *et al.*, 2008). Thus it is conceivable that these molecular events explain the underlying molecular basis of ChiLs role in the tissue remodelling (Johansen *et al.*, 1997). Since heparan sulphate is negatively charged molecule, the binding protein is expected to possess strong positively charged residues at the binding site (Esko and Selleck, 2002). On ChiLs this cluster is present separately to the central ligand binding groove which implies that central characteristic cleft of humans ChiLs may not be involved in their cellular proliferation and tissue remodelling function. Further supporting this notion is the involvement of BP40 in tissue remodelling in mammary gland tissues in members of family bovidae (Srivastava *et al.*, 2007). BP40 is a bovidae specific chilectin and despite the close relationship with ChiL1 it has a distinct central ligand binding cleft to CHIL1 but identitical potential heparin binding sites. Moreover, BP40 is unable to bind with a chitin tetramere (Kumar *et al*., 2007) and protein binding propensity has also been proposed for it (Mohanty *et al*., 2003). Taking both phylogenetic and structural data into account, it is conceivable that this probable heparin binding site on CHIL1 is in transition between neo to subfunctionalization (He and Zhang, 2005) since its origin in mammals. Comparing the same regions among different orthologues of CHIL1 may provide a signature for the ongoing evolution.

OVGP1 plays a role in fertilization and early embryonic development (Buhi, 2002). Cell surfaces of both mammalian ova and sperm are coated with different glycoproteins (Batova *et al*., 1998; Bauskin *et al*., 1999) and these are involved in establishing interactions between gametes during fertilization (Clark, 2013; Gupta *et al*., 2012). Therefore it is possible that OVGP1 coats mammalian gametes by binding with their surface glycoproteins via the carbohydrate moieties. Indeed co-incubation of human sperm (Boatman and Magnoni, 1995) and ovum (Martus *et al*., 1998) with OVGP1 increases sperm binding and penetration. Moreover, the presence of a mucin-like highly glycosylated C-terminal tail in OVGP1 further contributes to sperm viability (Satoh *et al*., 1995). This may suggests that OVGP1 along with its tail may have evolved for its new environment and function. However, it is not yet understood how the interaction of OVGP1 with gametes and zygote biologically affect the fertilization and early embryonic development. Moreover, studies conducted using Ovgp1 null mice (-/-) suggest that it is not essential for fertilization (Araki *et al*., 2003).

CHIT1 and CHIA share a common progenitor and are known to bind and hydrolyze chitin oligomers (Fusetti *et al*., 2002; Olland *et al*., 2009), but they are expressed in different tissues. Human CHIT1 is expressed in macrophages while CHIA largely restricted to the stomach. The exochitinase, CTBS, cavity is noticeably wide in comparison to other GH18 active chitinases, which may reflect its different mechanism of chitinase activity. However, the ligand binding specificity and functions of human and other mammalian CTBS, is not known. Among all the compared chilectins, CHID1 has the smallest central cavity volume, however it does have larger cavities separate from the central cleft; nevertheless binding with the carbohydrate moieties (preferably monomeric) and lipopolysaccharides with less selectivity has been reported for the protein. It has also been proposed that this binding between CHID1 and lipopolysaccharide may neutralize the endotoxin of the invading bacterial pathogens (Meng *et al*., 2010).

Finally, due to an irregularly shaped central cleft, interaction with oligomeric carbohydrates may not be feasible for certain ChiLs (Chil3, Chil4, Chil6, BP40, CHID1 and OVGP1). However, binding with shorter carbohydrate moieties may be possible. Hyaluronic acid (HA) serves as a structural analogue of invertebrate chitin in vertebrates and its synthesis requires short chito-oligosacchrides (Semino *et al*., 1996; Meyer *et al*., 1996). Additionally, the biological function of HA in tissue remodelling, inflammation and embryogenesis (Lee *et al*., 2000) over-laps with some of the functions proposed or known for Chil3 (Nio *et al*., 2004), BP40 (Srivastava *et al*., 2007), CHID1 (Meng *et al*., 2010) and OVGP1 (Buhi *et al*., 2002). Therefore it is tempting to speculate that these proteins undertake their tasks due to their involvement in the HA synthesis.

No significant sequential or structural differences have been observed between the compared orthologues of human and mouse GH18 proteins, suggesting functional similarity between the orthologous proteins. However, several of the orthologues show different tissue specific expression between the species. For example in mouse, Chia inhibits chitin induced inflammation and contributes to Th2 mediated adaptive immunity in the lungs (Zhu *et al*., 2004; Reese *et al*., 2007). However, unlike mouse, in humans CHIT1 not CHIA is the major chitinase in the lungs (Boot *et al*., 2005).

Taken together, the expansion and diversification of GH18 proteins in vertebrates is followed by acquiring tissue specificity and adopting subtle structural variations. This in turn may allow the proteins to optimize additional functions from the ancestral protein or result in a change in function. This evolution and potential change in the ligand binding specificity was paralleled with the extensive anatomical and physiological changes in vertebrates, especially in the mammalian immune system and reproductive system. Given the accounts of involvement of chitinases and ChiLs in the immune response it is reasonable to assume that change in the ligand specificity may have played some role for chitinase and chitinase like protein to adopt new role(s) in adaptive immune physiology. This study does not identify the natural ligands of the chitinases and ChiLs and previous experimental studies in this regard are inconclusive. However, it demonstrates that different chitinases and ChiLs may have different ligand specificity and points towards possible candidates. Further studies, like molecular docking, co-immuno precipitation, affinity purification and co crystallization with potential ligands, will provide more insight in this regard.

Chapter 7. Building Bridges

7. Building Bridges

Without the current advances in bioinformatics, many new fields of study such as “***Systems Biology***” and “***Comparative Genomics***” would largely remain a dream. In view of the 2013 Nobel Prize in chemistry going to computational chemistry, the importance of bioinformatics has clearly been recognized. This thesis deals with the application of some of the many bioinformatics tools to understand the structure-function relationship of the EBV encoded protein, EBNA1, and the evolution of two gene families, USPs and GH18.

Reliability of Protein Molecular Modelling

Protein-protein interactions are one of the hallmarks of biological complexity in living organisms. Molecular modelling of EBNA1 not only provides insights into the structural evolution of the molecule in different LCVs but also provide an indication of how EBNA1 might bind to multiple partners. However, the efficiency and consistency of prediction made by protein modelling tools varies considerably. Protein homology modelling algorithms have become significantly refined and have been more consistent in their predictions in recent years. Protein structural models showing an RMSD value of <2.0Å compared with the template, typically fulfil high structural requirements (including dihedral angle ratio, plausible angles, thermodynamic stability etc) and can be reliably used for molecular docking studies to explore ligand specificities and virtual drug screening (Ekins *et al*., 2007; Zhang et al., 2009).

Unlike homology modelling, the *ab initio* and iterative threading alignment methodologies for protein modelling show inconsistencies in their predictions. Discrepancies in the structural prediction are primarily due to the acquisition of evolutionary unrelated templates with marginal and/or fragmented sequence similarity with the query sequence, as seen in the predictions made by I-TASSER and MOE for the N-terminal half of EBNA1. Under these conditions, the RMSD values between the model and templates are irrelevant because of the possible misorientiation of loops and tails. However, the core region of the model may be correct. Despite this, approaches such as TM score can be used for evaluating models constructed in the absence suitable templates. TM scoring is mainly based on large distance between atoms of template(s) and models thus it is more sensitive to global topology than local structural errors. Therefore, the N-terminal half of EBNA1 and C-terminal tail of selected GH18 proteins, which were modelled using non homologous proteins, may contain local errors; however their TM values are within acceptable limits (3-4). Both TM scoring and Ramachandran Plot analysis suggest that EBNA1 models could be used to predict the global conformation of the molecules. Moreover due to the structural conservation between proteins sharing the same domains, the models could be used to identify domain boundaries and family and superfamily assignment (Malmstrom *et al*., 2007; Zhang *et al*., 2006). The presence of intrinsically unstructured regions (as observed in EBNA1) in a protein sequence may also contribute to the uncertainty in the prediction and spatial placement of loops and coils (Liu *et al*., 2009). The accurate prediction of loops is still a problem for computational modelling of the protein structures (Roy *et al*., 2010). Given that protein-protein interactions are dynamic and often involve conformational changes in the binding partners, such ambiguities limit the subsequent investigation using molecular docking (Goh *et al*., 2004; Zacharias, 2010).

Pitfalls of Phylogenomic Studies

Phylogenomic analyses are an important component of bioinformatic studies. The application of such studies ranges from unravelling the evolutionary history of a gene and of a species (Bininda-Emonds *et al*., 2007; Yang, 2013) to develop an understanding of molecular networks (Soyer and Malley, 2013) and even drug designing (Brown and Auger, 2011; Wang *et al*. 2013). In this thesis, phylogenetic relationships and the underlying evolutionary mechanisms of two gene families, USPs and GH18, have been investigated. All phylogenetic analyses are statistical in nature (Kumar *et al*., 2011) and as a result the inference of the trees depends on sample size, methods employed for their reconstruction and the numerical values in support. At present, maximum likelihood and bayesian posterior probability are the two most frequently adopted methodologies to infer the phylogenetic relationship between genes, species and gene families. However neither of these methods is completely perfect and can be affected by sequence composition and sample size (Yang and Rannala, 2012).

For practical convenience, sequences from representative species of the important speciation events were analysed to infer the evolutionary history of the very distantly related and extensively diversified USP gene family. Conversely, GH18 genes in vertebrates started to diversify with the origin of vertebrates and later with the emergence of mammals and rodents, therefore all available vertebrate sequences were included in the phylogenetic analysis of the GH18 gene family. Subject to the variable selection pressure and historical extent of divergence, occasionally sequences at the roots of clades could be difficult to infer with reasonable certitude potentially due to difficulty in the alignment. Additionally lacking in the sequence data, post speciation gene death can also lead to the misplacement of clades with respect to speciation events as seen in paralogous group 7 of USPs. These issues can be overcome or the tree approximated, by considering the biological context.

Future perspectives

Typically docking programmes use scoring criteria, based on desolvation energy, hydrophobicity and electrostatic residues of challenged residues, to predict orientation of partner molecules during interaction (Ritchi, 2008; Tovchigrechko and Vasker, 2006). However, the solution predicted using these scoring criteria are still a subject of debate. Incorporation of other bioinformatic and empirical analysis such as sequence conservation and protein interaction data retrieved from peptide array and other protein interaction assays could improve the selection of a correct native conformation of interacting protein molecules.

At present, a maximum length of 25 amino acids for each peptide could be spotted on an array, since the coupling efficiency drops after 12 amino acids. Given the possibility of structural properties of peptide (as observed by probing anti EBNA1 antibodies on EBNA1 array), attempts are required to spot the full structural and/or functional domain (as marked by sequence conservation and structural modelling) sequence on a single spot. This may rectify the potential structural variabilities between the spotted peptides and in turn result in improved consistencies of the peptide array data.

In this study the ClustalX programme was used to construct multiple sequence alignment. Nevertheless other alignment tools such as T-coffee (slower but relatively more accurate than ClustalX) could be used to achieve more accuracy in the alignment (Notredame *et al*., 2000). In addition, phylogeny aware sequence alignment tools could be explored for example PRANK (Loytynoja and Goldman, 2008), PAGAN (Loytynoja *et al*., 2012) and ProGraphMSA (Szalkowski, 2012) to improve the sequence alignments (most of these programmes are either computationally extensive or not available at the onset of the present study). Using these alignment tools may in turn improve the tree construction. Presently, MEGAv5.2 attempts to resolve the ambiguous branch positioning of the taxa in tree with some success by invoking a branch swap filter. Branch swap filter improves the stringency of the tree with respect to branch length and likelihood by altering the order of branches. Other parallel analyses, like evolutionary distance estimation, comparison of genomic synteny and protein domain architecture, provide important biological information to resolve ambiguities in the phylogenetic analysis. Additionally, the use of other efficient alternative methodologies for phylogenetic tree reconstruction (such as Bayesian based phylogeny) could also be informative in developing a holistic and more general conclusion (Anisimova *et al*., 2013). Finally combining the phylogeny with the structural features of molecules, as undertaken for EBNA1, USPs and GH18 family proteins, is also useful to extend the phylogenetic observations to functional inference for the genes.

A database called TimeTree has been established which provides the time line of speciation events during the course of eukaryotic evolution based on paleontological and phylogenetic information (Hedges *et al*., 2006). Recently, another database (BioName) has been developed to provide the evolutionary history of taxa based on text minining (Page, 2013). It would be of great value to develop a data base for the phylogenetic trees of genes and gene families where researchers can deposit their trees for comparison with other trees developed in other relevant studies and with the protein family trees present in different databases. Subsequently an algorithm could be designed to develop the consensus evolutionary relationship between orthologues and paralogues. This in turn will provide useful and readily available evolutionary information about different genes and gene families to users, especially to those who are working in other fields of life sciences where the phylogenomic information of molecules is relevant, but not the major objective of work.

As extensive genome sequencing projects are underway and new advances in computer technologies and bioinformatics are regularly reported, the present study is by no means complete. However, it provides new insights for the experimental and computational biologists to verify and further expand upon, which will eventually complete the bridges in our current understanding of structural, functional and evolutionary aspects of these genes and proteins.

As **Carl Sagan** famously put in his last interview on 27th May 1996.

"Science is more than a body of knowledge. It is a way of thinking; a way of skeptically interrogating the universe with a fine understanding of human fallibility"

8. Bibliography:

1. Adams, A., Lindahl, T., and Klein, G. (1973). Linear association between cellular DNA and Epstein-Barr virus DNA in a human lymphoblastoid cell line. Proc Natl Acad Sci U S A 70(10), 2888-92.
2. Agarwal, G., Rajavel, M., Gopal, B., and Srinivasan, N. (2009). Structure-based phylogeny as a diagnostic for functional characterization of proteins with a cupin fold. PLoS One 4(5), e5736.
3. Alexopoulou, A. N., Multhaupt, H. A., and Couchman, J. R. (2007). Syndecans in wound healing, inflammation and vascular biology. Int J Biochem Cell Biol 39(3), 505-28.
4. Allday, M. J., and Farrell, P. J. (1994). Epstein-Barr virus nuclear antigen EBNA3C/6 expression maintains the level of latent membrane protein 1 in G1-arrested cells. J Virol 68(6), 3491-8.
5. Altmann, M., Pich, D., Ruiss, R., Wang, J. D., Sugden, B., and Hammerschmidt, W. (2006). Transcriptional activation by EBV nuclear antigen 1 is essential for the expression of EBV's transforming genes. Proceedings of the National Academy of Sciences of the United States of America 103(38), 14188-14193.
6. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17), 3389-402.
7. Alvarez-Venegas, R., and Avramova, Z. (2012). Evolution of the PWWP-domain encoding genes in the plant and animal lineages. BMC Evol Biol 12, 101.
8. Ambinder, R. F., Shah, W. A., Rawlins, D. R., Hayward, G. S., and Hayward, S. D. (1990). Definition of the sequence requirements for binding of the EBNA-1 protein to its palindromic target sites in Epstein-Barr virus DNA. J Virol 64(5), 2369-79.
9. Amerik, A. Y., and Hochstrasser, M. (2004). Mechanism and function of deubiquitinating enzymes. Biochim Biophys Acta 1695(1-3), 189-207.
10. Amon, W., and Farrell, P. J. (2005). Reactivation of Epstein-Barr virus from latency. Rev Med Virol 15(3), 149-56.
11. Amores, A., Force, A., Yan, Y. L., Joly, L., Amemiya, C., Fritz, A., Ho, R. K., Langeland, J., Prince, V., Wang, Y. L., Westerfield, M., Ekker, M., and Postlethwait, J. H. (1998). Zebrafish hox clusters and vertebrate genome evolution. Science 282(5394), 1711-4.
12. Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N. J., Habel, U., Schneider, F., and Zilles, K. (2005). Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. Anat Embryol (Berl) 210(5-6), 343-52.
13. Anisimova, M., Liberles, D. A., Philippe, H., Provan, J., Pupko, T., and von Haeseler, A. (2013). State-of the art methodologies dictate new standards for phylogenetic analysis. BMC Evol Biol 13, 161.
14. Apcher, S., Komarova, A., Daskalogianni, C., Yin, Y., Malbert-Colas, L., and Fahraeus, R. (2009). mRNA translation regulation by the Gly-Ala repeat of Epstein-Barr virus nuclear antigen 1. J Virol 83(3), 1289-98.
15. Arakane, Y., and Muthukrishnan, S. (2010). Insect chitinase and chitinase-like proteins. Cell Mol Life Sci 67(2), 201-16.
16. Araki, Y., Nohara, M., Yoshida-Komiya, H., Kuramochi, T., Ito, M., Hoshi, H., Shinkai, Y., and Sendai, Y. (2003). Effect of a null mutation of the oviduct-specific glycoprotein gene on mouse fertilization. Biochem J 374(Pt 2), 551-7.
17. Aras, S., Singh, G., Johnston, K., Foster, T., and Aiyar, A. (2009). Zinc coordination is required for and regulates transcription activation by Epstein-Barr nuclear antigen 1. PLoS Pathog 5(6), e1000469.

18. Armengol, L., Marques-Bonet, T., Cheung, J., Khaja, R., Gonzalez, J. R., Scherer, S. W., Navarro, A., and Estivill, X. (2005). Murine segmental duplications are hot spots for chromosome and gene evolution. Genomics 86(6), 692-700.
19. Atanasiu, C., Deng, Z., Wiedmer, A., Norseen, J., and Lieberman, P. M. (2006). ORC binding to TRF2 stimulates OriP replication. EMBO Rep 7(7), 716-21.
20. Avery, S. V. (2011). Molecular targets of oxidative stress. Biochem J 434(2), 201-10.
21. Avvakumov, G. V., Walker, J. R., Xue, S., Finerty, P. J., Jr., Mackenzie, F., Newman, E. M., and Dhe-Paganon, S. (2006). Amino-terminal dimerization, NRDP1-rhodanese interaction, and inhibited catalytic domain conformation of the ubiquitin-specific protease 8 (USP8). J Biol Chem 281(49), 38061-70.
22. Babu, M. M., van der Lee, R., de Groot, N. S., and Gsponer, J. (2011). Intrinsically disordered proteins: regulation and disease. Curr Opin Struct Biol 21(3), 432-40.
23. Baer, R., Bankier, A. T., Biggin, M. D., Deininger, P. L., Farrell, P. J., Gibson, T. J., Hatfull, G., Hudson, G. S., Satchwell, S. C., Seguin, C., and et al. (1984). DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 310(5974), 207-11.
24. Baillie, G. S., Adams, D. R., Bhari, N., Houslay, T. M., Vadrevu, S., Meng, D., Li, X., Dunlop, A., Milligan, G., Bolger, G. B., Klussmann, E., and Houslay, M. D. (2007). Mapping binding sites for the PDE4D5 cAMP-specific phosphodiesterase to the N- and C-domains of beta-arrestin using spot-immobilized peptide arrays. Biochem J 404(1), 71-80.
25. Baker, N. E. (1987). Molecular cloning of sequences from wingless, a segment polarity gene in Drosophila: the spatial distribution of a transcript in embryos. EMBO J 6(6), 1765-73.
26. Barone, R., Simpore, J., Malaguarnera, L., Pignatelli, S., and Musumeci, S. (2003). Plasma chitotriosidase activity in acute Plasmodium falciparum malaria. Clin Chim Acta 331(1-2), 79-85.
27. Bashaw, J. M., and Yates, J. L. (2001). Replication from oriP of Epstein-Barr virus requires exact spacing of two bound dimers of EBNA1 which bend DNA. J Virol 75(22), 10603-11.
28. Bassermann, F., Frescas, D., Guardavaccaro, D., Busino, L., Peschiaroli, A., and Pagano, M. (2008). The Cdc14B-Cdh1-Plk1 axis controls the G2 DNA-damage-response checkpoint. Cell 134(2), 256-67.
29. Basu, M. K., Carmel, L., Rogozin, I. B., and Koonin, E. V. (2008). Evolution of protein domain promiscuity in eukaryotes. Genome Res 18(3), 449-61.
30. Batova, I. N., Ivanova, M. D., Mollova, M. V., and Kyurkchiev, S. D. (1998). Human sperm surface glycoprotein involved in sperm-zona pellucida interaction. Int J Androl 21(3), 141-53.
31. Battey, J. N., Kopp, J., Bordoli, L., Read, R. J., Clarke, N. D., and Schwede, T. (2007). Automated server predictions in CASP7. Proteins 69 Suppl 8, 68-82.
32. Baud, V., and Karin, M. (2009). Is NF-kappaB a good target for cancer therapy? Hopes and pitfalls. Nat Rev Drug Discov 8(1), 33-40.
33. Bauskin, A. R., Franken, D. R., Eberspaecher, U., and Donner, P. (1999). Characterization of human zona pellucida glycoproteins. Mol Hum Reprod 5(6), 534-40.
34. Bekpen, C., Tastekin, I., Siswara, P., Akdis, C. A., and Eichler, E. E. (2012). Primate segmental duplication creates novel promoters for the LRRC37 gene family within the 17q21.31 inversion polymorphism region. Genome Res 22(6), 1050-8.
35. Benkert, P., Kunzli, M., and Schwede, T. (2009). QMEAN server for protein model quality estimation. Nucleic Acids Res 37(Web Server issue), W510-4.

36. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Res 28(1), 235-42.

37. Bernardi, R., and Pandolfi, P. P. (2007). Structure, dynamics and functions of promyelocytic leukaemia nuclear bodies. Nat Rev Mol Cell Biol 8(12), 1006-16.

38. Berthouze, M., Venkataramanan, V., Li, Y., and Shenoy, S. K. (2009). The deubiquitinases USP33 and USP20 coordinate beta2 adrenergic receptor recycling and resensitization. EMBO J 28(12), 1684-96.

39. Bett, J. S., Ibrahim, A. F., Garg, A. K., Kelly, V., Pedrioli, P., Rocha, S., and Hay, R. T. (2013). The P-body component USP52/PAN2 is a novel regulator of HIF1A mRNA stability. Biochem J.

40. Bidartondo, M. I. (2008). Preserving accuracy in GenBank. Science 319(5870), 1616.

41. Biggar, R. J., Johansen, J. S., Smedby, K. E., Rostgaard, K., Chang, E. T., Adami, H. O., Glimelius, B., Molin, D., Hamilton-Dutoit, S., Melbye, M., and Hjalgrim, H. (2008). Serum YKL-40 and interleukin 6 levels in Hodgkin lymphoma. Clin Cancer Res 14(21), 6974-8.

42. Bignell, G. R., Warren, W., Seal, S., Takahashi, M., Rapley, E., Barfoot, R., Green, H., Brown, C., Biggs, P. J., Lakhani, S. R., Jones, C., Hansen, J., Blair, E., Hofmann, B., Siebert, R., Turner, G., Evans, D. G., Schrander-Stumpel, C., Beemer, F. A., van Den Ouweland, A., Halley, D., Delpech, B., Cleveland, M. G., Leigh, I., Leisti, J., and Rasmussen, S. (2000). Identification of the familial cylindromatosis tumour-suppressor gene. Nat Genet 25(2), 160-5.

43. Bininda-Emonds, O. R., Cardillo, M., Jones, K. E., MacPhee, R. D., Beck, R. M., Grenyer, R., Price, S. A., Vos, R. A., Gittleman, J. L., and Purvis, A. (2007). The delayed rise of present-day mammals. Nature 446(7135), 507-12.

44. Birdsey, G. M., Dryden, N. H., Shah, A. V., Hannah, R., Hall, M. D., Haskard, D. O., Parsons, M., Mason, J. C., Zvelebil, M., Gottgens, B., Ridley, A. J., and Randi, A. M. (2012). The transcription factor Erg regulates expression of histone deacetylase 6 and multiple pathways involved in endothelial cell migration and angiogenesis. Blood 119(3), 894-903.

45. Blake, N. W., Moghaddam, A., Rao, P., Kaur, A., Glickman, R., Cho, Y. G., Marchini, A., Haigh, T., Johnson, R. P., Rickinson, A. B., and Wang, F. (1999). Inhibition of antigen presentation by the glycine/alanine repeat domain is not conserved in simian homologues of Epstein-Barr virus nuclear antigen 1. J Virol 73(9), 7381-9.

46. Blum, K. A., Lozanski, G., and Byrd, J. C. (2004). Adult Burkitt leukemia and lymphoma. Blood 104(10), 3009-20.

47. Boatman, D. E., and Magnoni, G. E. (1995). Identification of a sperm penetration factor in the oviduct of the golden hamster. Biol Reprod 52(1), 199-207.

48. Bochkarev, A., Barwell, J. A., Pfuetzner, R. A., Bochkareva, E., Frappier, L., and Edwards, A. M. (1996). Crystal structure of the DNA-binding domain of the Epstein-Barr virus origin-binding protein, EBNA1, bound to DNA. Cell 84(5), 791-800.

49. Bochkarev, A., Barwell, J. A., Pfuetzner, R. A., Furey, W., Jr., Edwards, A. M., and Frappier, L. (1995). Crystal structure of the DNA-binding domain of the Epstein-Barr virus origin-binding protein EBNA 1. Cell 83(1), 39-46.

50. Boot, R. G., Blommaart, E. F., Swart, E., Ghauharali-van der Vlugt, K., Bijl, N., Moe, C., Place, A., and Aerts, J. M. (2001). Identification of a novel acidic mammalian chitinase distinct from chitotriosidase. J Biol Chem 276(9), 6770-8.

51. Boot, R. G., Bussink, A. P., Verhoek, M., de Boer, P. A., Moorman, A. F., and Aerts, J. M. (2005). Marked differences in tissue-specific expression of chitinases in mouse and man. J Histochem Cytochem 53(10), 1283-92.

52. Boot, R. G., Renkema, G. H., Verhoek, M., Strijland, A., Bliek, J., de Meulemeester, T. M., Mannens, M. M., and Aerts, J. M. (1998). The human chitotriosidase gene. Nature of inherited enzyme deficiency. J Biol Chem 273(40), 25680-5.
53. Boots, A. M., Hubers, H., Kouwijzer, M., den Hoed-van Zandbrink, L., Westrek-Esselink, B. M., van Doorn, C., Stenger, R., Bos, E. S., van Lierop, M. J., Verheijden, G. F., Timmers, C. M., and van Staveren, C. J. (2007). Identification of an altered peptide ligand based on the endogenously presented, rheumatoid arthritis-associated, human cartilage glycoprotein-39(263-275) epitope: an MHC anchor variant peptide for immune modulation. Arthritis Res Ther 9(4), R71.
54. Bordo, D., and Bork, P. (2002). The rhodanese/Cdc25 phosphatase superfamily. Sequence-structure-function relations. EMBO Rep 3(8), 741-6.
55. Bordo, D., Deriu, D., Colnaghi, R., Carpen, A., Pagani, S., and Bolognesi, M. (2000). The crystal structure of a sulfurtransferase from Azotobacter vinelandii highlights the evolutionary relationship between the rhodanese and phosphatase enzyme families. J Mol Biol 298(4), 691-704.
56. Bork, P. (1991). Shuffled domains in extracellular proteins. FEBS Lett 286(1-2), 47-54.
57. Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72, 248-54.
58. Brameld, K. A., and Goddard, W. A., 3rd (1998). The role of enzyme distortion in the single displacement mechanism of family 19 chitinases. Proc Natl Acad Sci U S A 95(8), 4276-81.
59. Brown, J. R., and Auger, K. R. (2011). Phylogenomics of phosphoinositide lipid kinases: perspectives on the evolution of second messenger signaling and drug discovery. BMC Evol Biol 11, 4.
60. Buckley, T. R., and Cunningham, C. W. (2002). The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. Mol Biol Evol 19(4), 394-405.
61. Buhi, W. C. (2002). Characterization and biological roles of oviduct-specific, oestrogen-dependent glycoprotein. Reproduction 123(3), 355-62.
62. Bukhari, S. A., and Caetano-Anolles, G. (2013). Origin and evolution of protein fold designs inferred from phylogenomic analysis of CATH domain structures in proteomes. PLoS Comput Biol 9(3), e1003009.
63. Burke, A. P., Yen, T. S., Shekitka, K. M., and Sobin, L. H. (1990). Lymphoepithelial carcinoma of the stomach with Epstein-Barr virus demonstrated by polymerase chain reaction. Mod Pathol 3(3), 377-80.
64. Burrows, J. F., Scott, C. J., and Johnston, J. A. (2010). The DUB/USP17 deubiquitinating enzymes: a gene family within a tandemly repeated sequence, is also embedded within the copy number variable beta-defensin cluster. BMC Genomics 11, 250.
65. Bussink, A. P., Speijer, D., Aerts, J. M., and Boot, R. G. (2007). Evolution of mammalian chitinase(-like) members of family 18 glycosyl hydrolases. Genetics 177(2), 959-70.
66. Bussink, A. P., van Eijk, M., Renkema, G. H., Aerts, J. M., and Boot, R. G. (2006). The biology of the Gaucher cell: the cradle of human chitinases. Int Rev Cytol 252, 71-128.
67. Calderone, T. L., Stevens, R. D., and Oas, T. G. (1996). High-level misincorporation of lysine for arginine at AGA codons in a fusion protein expressed in Escherichia coli. J Mol Biol 262(4), 407-12.

68. Caldwell, R. G., Wilson, J. B., Anderson, S. J., and Longnecker, R. (1998). Epstein-Barr virus LMP2A drives B cell development and survival in the absence of normal B cell receptor signals. Immunity 9(3), 405-11.
69. Canaan, A., Haviv, I., Urban, A. E., Schulz, V. P., Hartman, S., Zhang, Z., Palejev, D., Deisseroth, A. B., Lacy, J., Snyder, M., Gerstein, M., and Weissman, S. M. (2009). EBNA1 regulates cellular gene expression by binding cellular promoters. Proc Natl Acad Sci U S A 106(52), 22421-6.
70. Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37(Database issue), D233-8.
71. Cao, J. Y., Mansouri, S., and Frappier, L. (2012). Changes in the Nasopharyngeal Carcinoma Nuclear Proteome Induced by the EBNA1 Protein of Epstein-Barr Virus Reveal Potential Roles for EBNA1 in Metastasis and Oxidative Stress Responses. Journal of Virology 86(1), 382-394.
72. Casey, G., Neville, P. J., Liu, X., Plummer, S. J., Cicek, M. S., Krumroy, L. M., Curran, A. P., McGreevy, M. R., Catalona, W. J., Klein, E. A., and Witte, J. S. (2006). Podocalyxin variants and risk of prostate cancer and tumor aggressiveness. Hum Mol Genet 15(5), 735-41.
73. Catic, A., Fiebiger, E., Korbel, G. A., Blom, D., Galardy, P. J., and Ploegh, H. L. (2007). Screen for ISG15-crossreactive deubiquitinases. PLoS One 2(7), e679.
74. Chang, K. L., Chen, Y. Y., Shibata, D., and Weiss, L. M. (1992). Description of an in situ hybridization methodology for detection of Epstein-Barr virus RNA in paraffin-embedded tissues, with a survey of normal and neoplastic tissues. Diagn Mol Pathol 1(4), 246-55.
75. Chaudhuri, B., Xu, H., Todorov, I., Dutta, A., and Yates, J. L. (2001). Human DNA replication initiation factors, ORC and MCM, associate with oriP of Epstein-Barr virus. Proc Natl Acad Sci U S A 98(18), 10085-9.
76. Chen, R., Mintseris, J., Janin, J., and Weng, Z. (2003). A protein-protein docking benchmark. Proteins 52(1), 88-91.
77. Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66(Pt 1), 12-21.
78. Cheng, T. C., Hsieh, S. S., Hsu, W. L., Chen, Y. F., Ho, H. H., and Sheu, L. F. (2010). Expression of Epstein-Barr nuclear antigen 1 in gastric carcinoma cells is associated with enhanced tumorigenicity and reduced cisplatin sensitivity. International Journal of Oncology 36(1), 151-160.
79. Chomczynski, P., and Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 162(1), 156-9.
80. Chothia, C., and Gerstein, M. (1997). Protein evolution. How far can sequences diverge? Nature 385(6617), 579, 581.
81. Choudhuri, T., Murakami, M., Kaul, R., Sahu, S. K., Mohanty, S., Verma, S. C., Kumar, P., and Robertson, E. S. (2010). Nm23-H1 can induce cell cycle arrest and apoptosis in B cells. Cancer Biol Ther 9(12), 1065-78.
82. Clark, G. F. (2013). The role of carbohydrate recognition during human sperm-egg binding. Hum Reprod 28(3), 566-77.
83. Coffman, F. D. (2008). Chitinase 3-Like-1 (CHI3L1): a putative disease marker at the interface of proteomics and glycomics. Crit Rev Clin Lab Sci 45(6), 531-62.
84. Cohen, J. I. (2000). Epstein-Barr virus infection. N Engl J Med 343(7), 481-92.

85. Comeau, S. R., Gatchell, D. W., Vajda, S., and Camacho, C. J. (2004). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 20(1), 45-50.

86. Coppotelli, G., Mughal, N., and Masucci, M. G. (2013). The Gly-Ala repeat modulates the interaction of Epstein-Barr virus nuclear antigen-1 with cellular chromatin. Biochem Biophys Res Commun 431(4), 706-11.

87. Cozzetto, D., Kryshtafovych, A., Fidelis, K., Moult, J., Rost, B., and Tramontano, A. (2009). Evaluation of template-based models in CASP8 with standard measures. Proteins 77 Suppl 9, 18-28.

88. Cridland, J. A., Curley, E. Z., Wykes, M. N., Schroder, K., Sweet, M. J., Roberts, T. L., Ragan, M. A., Kassahn, K. S., and Stacey, K. J. (2012). The mammalian PYHIN gene family: phylogeny, evolution and expression. BMC Evol Biol 12, 140.

89. Cross, F. R., Buchler, N. E., and Skotheim, J. M. (2011). Evolution of networks and sequences in eukaryotic cell cycle control. Philos Trans R Soc Lond B Biol Sci 366(1584), 3532-44.

90. Dahiya, N., Tewari, R., and Hoondal, G. S. (2006). Biotechnological aspects of chitinolytic enzymes: a review. Appl Microbiol Biotechnol 71(6), 773-82.

91. Dawson, C. W., Tramountanis, G., Eliopoulos, A. G., and Young, L. S. (2003). Epstein-Barr virus latent membrane protein 1 (LMP1) activates the phosphatidylinositol 3-kinase/Akt pathway to promote cell survival and induce actin filament remodeling. J Biol Chem 278(6), 3694-704.

92. de Lichtenberg, U., Jensen, T. S., Brunak, S., Bork, P., and Jensen, L. J. (2007). Evolution of cell cycle control: same molecular machines, different regulation. Cell Cycle 6(15), 1819-25.

93. de The, G., Ablashi, D. V., Liabeuf, A., and Mourali, N. (1973). Nasopharyngeal carcinoma (NPC). VI. Presence of an EBV nuclear antigen in fresh tumour biopsies. Preliminary results. Biomedicine 19(8), 349-52.

94. Decaussin, G., Sbih-Lammali, F., de Turenne-Tessier, M., Bouguermouh, A., and Ooka, T. (2000). Expression of BARF1 gene encoded by Epstein-Barr virus in nasopharyngeal carcinoma biopsies. Cancer Res 60(19), 5584-8.

95. Degnan, J. H., and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. PLoS Genet 2(5), e68.

96. Degnan, J. H., and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24(6), 332-40.

97. Dehal, P., and Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. Plos Biology 3(10), e314.

98. Deleage, G., and Roux, B. (1987). An algorithm for protein secondary structure prediction based on class prediction. Protein Eng 1(4), 289-94.

99. DeLorenze, G. N., Munger, K. L., Lennette, E. T., Orentreich, N., Vogelman, J. H., and Ascherio, A. (2006). Epstein-Barr virus and multiple sclerosis: evidence of association from a prospective study with long-term follow-up. Arch Neurol 63(6), 839-44.

100. Demuth, J. P., and Hahn, M. W. (2009). The life and death of gene families. Bioessays 31(1), 29-39.

101. Deng, C. X., and Wang, R. H. (2003a). Roles of BRCA1 in DNA damage repair: a link between development and cancer. Hum Mol Genet 12 Spec No 1, R113-23.

102. Deng, Z., Atanasiu, C., Burg, J. S., Broccoli, D., and Lieberman, P. M. (2003b). Telomere repeat binding factors TRF1, TRF2, and hRAP1 modulate replication of Epstein-Barr virus OriP. J Virol 77(22), 11992-2001.

103. Deng, Z., Atanasiu, C., Zhao, K., Marmorstein, R., Sbodio, J. I., Chi, N. W., and Lieberman, P. M. (2005). Inhibition of Epstein-Barr virus OriP function by

tankyrase, a telomere-associated poly-ADP ribose polymerase that binds and modifies EBNA1. J Virol 79(8), 4640-50.

104. Deng, Z., Lezina, L., Chen, C. J., Shtivelband, S., So, W., and Lieberman, P. M. (2002). Telomeric proteins regulate episomal maintenance of Epstein-Barr virus origin of plasmid replication. Mol Cell 9(3), 493-503.

105. Dessimoz, C., and Gil, M. (2008). Covariance of maximum likelihood evolutionary distances between sequences aligned pairwise. BMC Evol Biol 8, 179.

106. Dhar, S. K., Yoshida, K., Machida, Y., Khaira, P., Chaudhuri, B., Wohlschlegel, J. A., Leffak, M., Yates, J., and Dutta, A. (2001). Replication from oriP of Epstein-Barr virus requires human ORC and is inhibited by geminin. Cell 106(3), 287-96.

107. Dhar, V., and Schildkraut, C. L. (1991). Role of EBNA-1 in arresting replication forks at the Epstein-Barr virus oriP family of tandem repeats. Mol Cell Biol 11(12), 6268-78.

108. Dheekollu, J., Deng, Z., Wiedmer, A., Weitzman, M. D., and Lieberman, P. M. (2007). A role for MRE11, NBS1, and recombination junctions in replication and stable maintenance of EBV episomes. PLoS One 2(12), e1257.

109. Dheekollu, J., and Lieberman, P. M. (2011). The replisome pausing factor Timeless is required for episomal maintenance of latent Epstein-Barr virus. J Virol 85(12), 5853-63.

110. d'Herouel, A. F., Birgersdotter, A., and Werner, M. (2010). FR-like EBNA1 binding repeats in the human genome. Virology 405(2), 524-9.

111. Dover, G. (1982). Molecular drive: a cohesive mode of species evolution. Nature 299(5879), 111-7.

112. Drag, M., Mikolajczyk, J., Bekes, M., Reyes-Turcu, F. E., Ellman, J. A., Wilkinson, K. D., and Salvesen, G. S. (2008). Positional-scanning fluorigenic substrate libraries reveal unexpected specificity determinants of DUBs (deubiquitinating enzymes). Biochem J 415(3), 367-75.

113. Dresang, L. R., Vereide, D. T., and Sugden, B. (2009). Identifying sites bound by Epstein-Barr virus nuclear antigen 1 (EBNA1) in the human genome: defining a position-weighted matrix to predict sites bound by EBNA1 in viral genomes. J Virol 83(7), 2930-40.

114. Drotar, M. E., Silva, S., Barone, E., Campbell, D., Tsimbouri, P., Jurvansu, J., Bhatia, P., Klein, G., and Wilson, J. B. (2003). Epstein-Barr virus nuclear antigen-1 and Myc cooperate in lymphomagenesis. International Journal of Cancer 106(3), 388-395.

115. Duellman, S. J., Thompson, K. L., Coon, J. J., and Burgess, R. R. (2009). Phosphorylation sites of Epstein-Barr virus EBNA1 regulate its function. J Gen Virol 90(Pt 9), 2251-9.

116. Dupont, S., Mamidi, A., Cordenonsi, M., Montagner, M., Zacchigna, L., Adorno, M., Martello, G., Stinchfield, M. J., Soligo, S., Morsut, L., Inui, M., Moro, S., Modena, N., Argenton, F., Newfeld, S. J., and Piccolo, S. (2009). FAM/USP9x, a deubiquitinating enzyme essential for TGFbeta signaling, controls Smad4 monoubiquitination. Cell 136(1), 123-35.

117. Dyson, P. J., and Farrell, P. J. (1985). Chromatin structure of Epstein-Barr virus. J Gen Virol 66 (Pt 9), 1931-40.

118. Edwards, R. H., Marquitz, A. R., and Raab-Traub, N. (2008). Epstein-Barr virus BART microRNAs are produced from a large intron prior to splicing. J Virol 82(18), 9094-106.

119. Ehlers, B., Spiess, K., Leendertz, F., Peeters, M., Boesch, C., Gatherer, D., and McGeoch, D. J. (2010). Lymphocryptovirus phylogeny and the origins of Epstein-Barr virus. J Gen Virol 91(Pt 3), 630-42.

120. Eichhorn, P. J., Rodon, L., Gonzalez-Junca, A., Dirac, A., Gili, M., Martinez-Saez, E., Aura, C., Barba, I., Peg, V., Prat, A., Cuartas, I., Jimenez, J., Garcia-Dorado, D.,

Sahuquillo, J., Bernards, R., Baselga, J., and Seoane, J. (2012). USP15 stabilizes TGF-beta receptor I and promotes oncogenesis through the activation of TGF-beta signaling in glioblastoma. Nat Med 18(3), 429-35.

121. Ekins, S., Mestres, J., and Testa, B. (2007). In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br J Pharmacol 152(1), 9-20.

122. Eliopoulos, A. G., Gallagher, N. J., Blake, S. M., Dawson, C. W., and Young, L. S. (1999). Activation of the p38 mitogen-activated protein kinase pathway by Epstein-Barr virus-encoded latent membrane protein 1 coregulates interleukin-6 and interleukin-8 production. J Biol Chem 274(23), 16085-96.

123. Eliopoulos, A. G., and Young, L. S. (2001). LMP1 structure and signal transduction. Semin Cancer Biol 11(6), 435-44.

124. Emanuel, B. S., and Shaikh, T. H. (2001). Segmental duplications: an 'expanding' role in genomic instability and disease. Nat Rev Genet 2(10), 791-800.

125. Epstein, M. A., Achong, B. G., and Barr, Y. M. (1964). Virus Particles in Cultured Lymphoblasts from Burkitt's Lymphoma. Lancet 1(7335), 702-3.

126. Escriva, H., Manzon, L., Youson, J., and Laudet, V. (2002). Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. Mol Biol Evol 19(9), 1440-50.

127. Esko, J. D., and Selleck, S. B. (2002). Order out of chaos: assembly of ligand binding sites in heparan sulfate. Annu Rev Biochem 71, 435-71.

128. Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U., and Sali, A. (2006). Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics Chapter 5, Unit 5 6.

129. Fan, Y. H., Yu, Y., Mao, R. F., Tan, X. J., Xu, G. F., Zhang, H., Lu, X. B., Fu, S. B., and Yang, J. (2011). USP4 targets TAK1 to downregulate TNFalpha-induced NF-kappaB activation. Cell Death and Differentiation 18(10), 1547-60.

130. Farrell, C. J., Lee, J. M., Shin, E. C., Cebrat, M., Cole, P. A., and Hayward, S. D. (2004). Inhibition of Epstein-Barr virus-induced growth proliferation by a nuclear antigen EBNA2-TAT peptide. Proc Natl Acad Sci U S A 101(13), 4625-30.

131. Farrell, P. J. (2005). Can plasma Epstein-Barr virus DNA levels be used to monitor nasopharyngeal carcinoma progression? Nat Clin Pract Oncol 2(1), 14-5.

132. Fechteler, T., Dengler, U., and Schomburg, D. (1995). Prediction of protein three-dimensional structures in insertion and deletion regions: a procedure for searching data bases of representative protein fragments using geometric scoring criteria. J Mol Biol 253(1), 114-31.

133. Fines BCaH, G. (2010). Chitinase and apparent digestibility of chitin in the digestive tract of juvenile cobia, Rachycentron canadum. Aquaculture 303, 34-39.

134. Fingeroth, J. D., Weis, J. J., Tedder, T. F., Strominger, J. L., Biro, P. A., and Fearon, D. T. (1984). Epstein-Barr virus receptor of human B lymphocytes is the C3d receptor CR2. Proc Natl Acad Sci U S A 81(14), 4510-4.

135. Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39(Web Server issue), W29-37.

136. Fiser, A., and Sali, A. (2003). ModLoop: automated modeling of loops in protein structures. Bioinformatics 19(18), 2500-1.

137. Flavell, J. R., Baumforth, K. R. N., Wood, V. H. J., Davies, G. L., Wei, W. B., Reynolds, G. M., Morgan, S., Boyce, A., Kelly, G. L., Young, L. S., and Murray, P. G. (2008). Down-regulation of the TGF-beta target gene, PTPRK, by the Epstein-Barr virus-encoded EBNA1 contributes to the growth and survival of Hodgkin lymphoma cells. Blood 111(1), 292-301.

138. Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Garcia-Giron, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kahari, A. K., Keenan, S.,

Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W. M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ritchie, G. R., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T. J., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., and Searle, S. M. (2013). Ensembl 2013. Nucleic Acids Res 41(Database issue), D48-55.

139. Forslund, K., Henricson, A., Hollich, V., and Sonnhammer, E. L. (2008). Domain tree-based analysis of protein architecture evolution. Mol Biol Evol 25(2), 254-64.

140. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41(Database issue), D808-15.

141. Frappier, L. (2012). EBNA1 and host factors in Epstein-Barr virus latent DNA replication. Curr Opin Virol 2(6), 733-9.

142. Frappier, L., and O'Donnell, M. (1991). Epstein-Barr nuclear antigen 1 mediates a DNA loop within the latent replication origin of Epstein-Barr virus. Proc Natl Acad Sci U S A 88(23), 10875-9.

143. Fruehling, S., and Longnecker, R. (1997). The immunoreceptor tyrosine-based activation motif of Epstein-Barr virus LMP2A is essential for blocking BCR-mediated signal transduction. Virology 235(2), 241-51.

144. Fu, T., Voo, K. S., and Wang, R. F. (2004). Critical role of EBNA1-specific CD4+ T cells in the control of mouse Burkitt lymphoma in vivo. J Clin Invest 114(4), 542-50.

145. Fujita, K., Shimomura, K., Yamamoto, K., Yamashita, T., and Suzuki, K. (2006). A chitinase structurally related to the glycoside hydrolase family 48 is indispensable for the hormonally induced diapause termination in a beetle. Biochem Biophys Res Commun 345(1), 502-7.

146. Fujiwara, K., Toda, H., and Ikeguchi, M. (2012). Dependence of alpha-helical and beta-sheet amino acid propensities on the overall protein fold type. BMC Struct Biol 12, 18.

147. Fukayama, M., Chong, J. M., and Uozaki, H. (2001). Pathology and molecular pathology of Epstein-Barr virus-associated gastric carcinoma. Curr Top Microbiol Immunol 258, 91-102.

148. Funkhouser, J. D., and Aronson, N. N., Jr. (2007). Chitinase family GH18: evolutionary insights from the genomic history of a diverse protein family. BMC Evol Biol 7, 96.

149. Fusetti, F., Pijning, T., Kalk, K. H., Bos, E., and Dijkstra, B. W. (2003). Crystal structure and carbohydrate-binding properties of the human cartilage glycoprotein-39. J Biol Chem 278(39), 37753-60.

150. Fusetti, F., von Moeller, H., Houston, D., Rozeboom, H. J., Dijkstra, B. W., Boot, R. G., Aerts, J. M., and van Aalten, D. M. (2002). Structure of human chitotriosidase. Implications for specific inhibitor design and function of mammalian chitinase-like lectins. J Biol Chem 277(28), 25537-44.

151. Futaki, S., Nakase, I., Suzuki, T., Youjun, Z., and Sugiura, Y. (2002). Translocation of branched-chain arginine peptides through cell membranes: flexibility in the spatial disposition of positive charges in membrane-permeable peptides. Biochemistry 41(25), 7925-30.

152. Gahn, T. A., and Sugden, B. (1995). An EBNA-1-dependent enhancer acts from a distance of 10 kilobase pairs to increase expression of the Epstein-Barr virus LMP gene. J Virol 69(4), 2633-6.

153. Gandhi, M. K., Tellam, J. T., and Khanna, R. (2004). Epstein-Barr virus-associated Hodgkin's lymphoma. Br J Haematol 125(3), 267-81.

154. Gelebart, P., Zak, Z., Anand, M., Belch, A., and Lai, R. (2012). Blockade of fatty acid synthase triggers significant apoptosis in mantle cell lymphoma. PLoS One 7(4), e33738.

155. Gerchman, S. E., Graziano, V., and Ramakrishnan, V. (1994). Expression of chicken linker histones in E. coli: sources of problems and methods for overcoming some of the difficulties. Protein Expr Purif 5(3), 242-51.

156. Gillis, W. Q., St John, J., Bowerman, B., and Schneider, S. Q. (2009). Whole genome duplications and expansion of the vertebrate GATA transcription factor gene family. BMC Evol Biol 9, 207.

157. Goh, C. S., Milburn, D., and Gerstein, M. (2004). Conformational changes associated with protein-protein interactions. Curr Opin Struct Biol 14(1), 104-9.

158. Golden, R. L. (1968). Infectious mononucleosis with Epstein-Barr virus antibodies in older ages. JAMA 205(8), 595.

159. Goldman, J. M., and Aisenberg, A. C. (1970). Incidence of antibody to EB virus, herpes simplex, and cytomegalovirus in Hodgkin's disease. Cancer 26(2), 327-31.

160. Gong, L., Kamitani, T., Millas, S., and Yeh, E. T. (2000). Identification of a novel isopeptidase with dual specificity for ubiquitin- and NEDD8-conjugated proteins. J Biol Chem 275(19), 14212-6.

161. Graner, E., Tang, D., Rossi, S., Baron, A., Migita, T., Weinstein, L. J., Lechpammer, M., Huesken, D., Zimmermann, J., Signoretti, S., and Loda, M. (2004). The isopeptidase USP2a regulates the stability of fatty acid synthase in prostate cancer. Cancer Cell 5(3), 253-61.

162. Graves, J. A., Koina, E., and Sankovic, N. (2006). How the gene content of human sex chromosomes evolved. Curr Opin Genet Dev 16(3), 219-24.

163. Greenspan, J. S., Greenspan, D., Lennette, E. T., Abrams, D. I., Conant, M. A., Petersen, V., and Freese, U. K. (1985). Replication of Epstein-Barr virus within the epithelial cells of oral "hairy" leukoplakia, an AIDS-associated lesion. N Engl J Med 313(25), 1564-71.

164. Grishin, N. V. (2001). Fold change in evolution of protein structures. J Struct Biol 134(2-3), 167-85.

165. Gross, H., Hennard, C., Masouris, I., Cassel, C., Barth, S., Stober-Grasser, U., Mamiani, A., Moritz, B., Ostareck, D., Ostareck-Lederer, A., Neuenkirchen, N., Fischer, U., Deng, W., Leonhardt, H., Noessner, E., Kremmer, E., and Grasser, F. A. (2012). Binding of the heterogeneous ribonucleoprotein K (hnRNP K) to the Epstein-Barr virus nuclear antigen 2 (EBNA2) enhances viral LMP2A expression. PLoS One 7(8), e42106.

166. Grossman, S. R., Johannsen, E., Tong, X., Yalamanchili, R., and Kieff, E. (1994). The Epstein-Barr virus nuclear antigen 2 transactivator is directed to response elements by the J kappa recombination signal binding protein. Proc Natl Acad Sci U S A 91(16), 7568-72.

167. Gruhne, B., Sompallae, R., Marescotti, D., Kamranvar, S. A., Gastaldello, S., and Masucci, M. G. (2009a). The Epstein-Barr virus nuclear antigen-1 promotes genomic instability via induction of reactive oxygen species. Proceedings of the National Academy of Sciences of the United States of America 106(7), 2313-2318.

168. Gruhne, B., Sompallae, R., and Masucci, M. G. (2009b). Three Epstein-Barr virus latency proteins independently promote genomic instability by inducing DNA damage, inhibiting DNA repair and inactivating cell cycle checkpoints. Oncogene 28(45), 3997-4008.

169. Gsponer, J., and Babu, M. M. (2009). The rules of disorder or why disorder rules. Prog Biophys Mol Biol 99(2-3), 94-103.

170. Guerriero, G. (2012). Putative chitin synthases from Branchiostoma floridae show extracellular matrix-related domains and mosaic structures. Genomics Proteomics Bioinformatics 10(4), 197-207.

171. Guo, T. W., Zhang, F. C., Yang, M. S., Gao, X. C., Bian, L., Duan, S. W., Zheng, Z. J., Gao, J. J., Wang, H., Li, R. L., Feng, G. Y., St Clair, D., and He, L. (2004). Positive association of the DIO2 (deiodinase type 2) gene with mental retardation in the iodine-deficient areas of China. J Med Genet 41(8), 585-90.

172. Gupta, S. K., Bhandari, B., Shrestha, A., Biswal, B. K., Palaniappan, C., Malhotra, S. S., and Gupta, N. (2012). Mammalian zona pellucida glycoproteins: structure and function during fertilization. Cell Tissue Res 349(3), 665-78.

173. Gutierrez, M. I., Bhatia, K., Barriga, F., Diez, B., Muriel, F. S., de Andreas, M. L., Epelman, S., Risueno, C., and Magrath, I. T. (1992). Molecular epidemiology of Burkitt's lymphoma from South America: differences in breakpoint location and Epstein-Barr virus association from tumors in other world regions. Blood 79(12), 3261-6.

174. Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. 41, 95-98.

175. Hammerschmidt, W., and Sugden, B. (1988). Identification and characterization of oriLyt, a lytic origin of DNA replication of Epstein-Barr virus. Cell 55(3), 427-33.

176. Hammond-Martel, I., Yu, H., and Affar el, B. (2012). Roles of ubiquitin signaling in transcription regulation. Cell Signal 24(2), 410-21.

177. Hannigan, A., Qureshi, A. M., Nixon, C., Tsimbouri, P. M., Jones, S., Philbey, A. W., and Wilson, J. B. (2011). Lymphocyte deficiency limits Epstein-Barr virus latent membrane protein 1 induced chronic inflammation and carcinogenic pathology in vivo. Mol Cancer 10(1), 11.

178. Hannigan, A., and Wilson, J. B. (2010). Evaluation of LMP1 of Epstein-Barr virus as a therapeutic target by its inhibition. Mol Cancer 9, 184.

179. Hearing, J. C., Lewis, A., and Levine, A. J. (1985). Structure of the Epstein-Barr virus nuclear antigen as probed with monoclonal antibodies. Virology 142(1), 215-20.

180. Hedges, S. B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22(23), 2971-2.

181. Henderson, S., Rowe, M., Gregory, C., Croom-Carter, D., Wang, F., Longnecker, R., Kieff, E., and Rickinson, A. (1991). Induction of bcl-2 expression by Epstein-Barr virus latent membrane protein 1 protects infected B cells from programmed cell death. Cell 65(7), 1107-15.

182. Henle, G., and Henle, W. (1976). Epstein-Barr virus-specific IgA serum antibodies as an outstanding feature of nasopharyngeal carcinoma. Int J Cancer 17(1), 1-7.

183. Hennessy, K., and Kieff, E. (1983). One of two Epstein-Barr virus nuclear antigens contains a glycine-alanine copolymer domain. Proc Natl Acad Sci U S A 80(18), 5665-9.

184. Herbert, A., and Rich, A. (1999). RNA processing and the evolution of eukaryotes. Nat Genet 21(3), 265-9.

185. Hochstrasser, M. (2009). Origin and function of ubiquitin-like proteins. Nature 458(7237), 422-9.

186. Hoeijmakers, J. H. (2009). DNA damage, aging, and cancer. N Engl J Med 361(15), 1475-85.

187. Holland, P. W., Garcia-Fernandez, J., Williams, N. A., and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. Dev Suppl, 125-33.

188. Holowaty, M. N., Zeghouf, M., Wu, H., Tellam, J., Athanasopoulos, V., Greenblatt, J., and Frappier, L. (2003). Protein profiling with Epstein-Barr nuclear antigen-1 reveals an interaction with the herpesvirus-associated ubiquitin-specific protease HAUSP/USP7. J Biol Chem 278(32), 29987-94.

189. Hong, M., Murai, Y., Kutsuna, T., Takahashi, H., Nomoto, K., Cheng, C. M., Ishizawa, S., Zhao, Q. L., Ogawa, R., Harmon, B. V., Tsuneyama, K., and Takano, Y. (2006). Suppression of Epstein-Barr nuclear antigen 1 (EBNA1) by RNA

interference inhibits proliferation of EBV-positive Burkitt's lymphoma cells. Journal of Cancer Research and Clinical Oncology 132(1), 1-8.

190. Houston, D. R., Recklies, A. D., Krupa, J. C., and van Aalten, D. M. (2003). Structure and ligand-induced conformational change of the 39-kDa glycoprotein from human articular chondrocytes. J Biol Chem 278(32), 30206-12.

191. Howe, J. G., and Shu, M. D. (1988). Isolation and characterization of the genes for two small RNAs of herpesvirus papio and their comparison with Epstein-Barr virus-encoded EBER RNAs. J Virol 62(8), 2790-8.

192. Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W., and Zimmermann, P. (2008). Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinformatics 2008, 420747.

193. Hu, M., Gu, L. C., Li, M. Y., Jeffrey, P. D., Gu, W., and Shi, Y. G. (2006). Structural basis of competitive recognition of p53 and MDM2 by HAUSP/USP7: Implications for the regulation of the p53-MDM2 pathway. Plos Biology 4(2), 228-239.

194. Hu, M., Li, P., Li, M., Li, W., Yao, T., Wu, J. W., Gu, W., Cohen, R. E., and Shi, Y. (2002). Crystal structure of a UBP-family deubiquitinating enzyme in isolation and in complex with ubiquitin aldehyde. Cell 111(7), 1041-54.

195. Hu, M., Li, P., Song, L., Jeffrey, P. D., Chenova, T. A., Wilkinson, K. D., Cohen, R. E., and Shi, Y. (2005). Structure and mechanisms of the proteasome-associated deubiquitinating enzyme USP14. EMBO J 24(21), 3747-56.

196. Huang, Q. S., Xie, X. L., Liang, G., Gong, F., Wang, Y., Wei, X. Q., Wang, Q., Ji, Z. L., and Chen, Q. X. (2012). The GH18 family of chitinases: their domain architectures, functions and evolutions. Glycobiology 22(1), 23-34.

197. Huang, T. T., Nijman, S. M., Mirchandani, K. D., Galardy, P. J., Cohn, M. A., Haas, W., Gygi, S. P., Ploegh, H. L., Bernards, R., and D'Andrea, A. D. (2006). Regulation of monoubiquitinated PCNA by DUB autocleavage. Nat Cell Biol 8(4), 339-47.

198. Huang, X., Summers, M. K., Pham, V., Lill, J. R., Liu, J., Lee, G., Kirkpatrick, D. S., Jackson, P. K., Fang, G., and Dixit, V. M. (2011). Deubiquitinase USP37 is activated by CDK2 to antagonize APC(CDH1) and promote S phase entry. Mol Cell 42(4), 511-23.

199. Huen, D. S., Henderson, S. A., Croom-Carter, D., and Rowe, M. (1995). The Epstein-Barr virus latent membrane protein-1 (LMP1) mediates activation of NF-kappa B and cell surface phenotype via two effector regions in its carboxy-terminal cytoplasmic domain. Oncogene 10(3), 549-60.

200. Hufton, A. L., Groth, D., Vingron, M., Lehrach, H., Poustka, A. J., and Panopoulou, G. (2008). Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. Genome Res 18(10), 1582-91.

201. Hume S, R. G., Feederele R, Delecluse H.J., Bousset K., Hammerschimdt W., Schepers A. (2003). The EBV nuclear antigen 1 (EBNA 1) enhances B cell immortalization several thousand fold. Proc. Natl. Acad. Sci. U. S. A. 100, 6.

202. Huminiecki, L., Goldovsky, L., Freilich, S., Moustakas, A., Ouzounis, C., and Heldin, C. H. (2009). Emergence, development and diversification of the TGF-beta signalling pathway within the animal kingdom. BMC Evol Biol 9, 28.

203. Humme, S., Reisbach, G., Feederle, R., Delecluse, H. J., Bousset, K., Hammerschmidt, W., and Schepers, A. (2003). The EBV nuclear antigen 1 (EBNA1) enhances B cell immortalization several thousandfold. Proc Natl Acad Sci U S A 100(19), 10989-94.

204. Hussain, M., and Wilson, J. B. (2013). New Paralogues and Revised Time Line in the Expansion of the Vertebrate GH18 Family. J Mol Evol.

205. Hussain, S., Zhang, Y., and Galardy, P. J. (2009). DUBs and cancer: the role of deubiquitinating enzymes as oncogenes, non-oncogenes and tumor suppressors. Cell Cycle 8(11), 1688-97.

206. Hutt-Fletcher, L. M. (2007). Epstein-Barr virus entry. J Virol 81(15), 7825-32.

207. Imai, S., Mamiya, T., Tsukada, A., Sakai, Y., Mouri, A., Nabeshima, T., and Ebihara, S. (2012). Ubiquitin-specific peptidase 46 (Usp46) regulates mouse immobile behavior in the tail suspension test through the GABAergic system. PLoS One 7(6), e39084.

208. Inbar, Y., Schneidman-Duhovny, D., Halperin, I., Oron, A., Nussinov, R., and Wolfson, H. J. (2005). Approaching the CAPRI challenge with an efficient geometry-based docking. Proteins 60(2), 217-23.

209. Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11(2), 97-108.

210. Jarrett, R. F., Stark, G. L., White, J., Angus, B., Alexander, F. E., Krajewski, A. S., Freeland, J., Taylor, G. M., and Taylor, P. R. (2005). Impact of tumor Epstein-Barr virus status on presenting features and outcome in age-defined subgroups of patients with classic Hodgkin lymphoma: a population-based study. Blood 106(7), 2444-51.

211. Jenson, H. B., Ench, Y., Zhang, Y., Gao, S. J., Arrand, J. R., and Mackett, M. (2002). Characterization of an Epstein-Barr virus-related gammaherpesvirus from common marmoset (Callithrix jacchus). J Gen Virol 83(Pt 7), 1621-33.

212. Johannsen, E., Luftig, M., Chase, M. R., Weicksel, S., Cahir-McFarland, E., Illanes, D., Sarracino, D., and Kieff, E. (2004). Proteins of purified Epstein-Barr virus. Proc Natl Acad Sci U S A 101(46), 16286-91.

213. Johansen, J. S., Moller, S., Price, P. A., Bendtsen, F., Junge, J., Garbarsch, C., and Henriksen, J. H. (1997). Plasma YKL-40: a new potential marker of fibrosis in patients with alcoholic cirrhosis? Scand J Gastroenterol 32(6), 582-90.

214. Johansson, M. U., Zoete, V., Michielin, O., and Guex, N. (2012). Defining and searching for structural motifs using DeepView/Swiss-PdbViewer. BMC Bioinformatics 13, 173.

215. Johnston, S. C., Larsen, C. N., Cook, W. J., Wilkinson, K. D., and Hill, C. P. (1997). Crystal structure of a deubiquitinating enzyme (human UCH-L3) at 1.8 A resolution. EMBO J 16(13), 3787-96.

216. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292(2), 195-202.

217. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8(3), 275-82.

218. Joo, H. Y., Zhai, L., Yang, C., Nie, S., Erdjument-Bromage, H., Tempst, P., Chang, C., and Wang, H. (2007). Regulation of cell cycle progression and gene expression by H2A deubiquitination. Nature 449(7165), 1068-72.

219. Jourdan, N., Jobart-Malfait, A., Dos Reis, G., Quignon, F., Piolot, T., Klein, C., Tramier, M., Coppey-Moisan, M., and Marechal, V. (2012). Live-cell imaging reveals multiple interactions between Epstein-Barr virus nuclear antigen 1 and cellular chromatin during interphase and mitosis. J Virol 86(9), 5314-29.

220. Kaiser, C., Laux, G., Eick, D., Jochner, N., Bornkamm, G. W., and Kempkes, B. (1999). The proto-oncogene c-myc is a direct target gene of Epstein-Barr virus nuclear antigen 2. J Virol 73(5), 4481-4.

221. Kamranvar, S. A., and Masucci, M. G. (2011). The Epstein-Barr virus nuclear antigen-1 promotes telomere dysfunction via induction of oxidative stress. Leukemia 25(6), 1017-1025.

222. Kane, J. F., Violand, B. N., Curran, D. F., Staten, N. R., Duffin, K. L., and Bogosian, G. (1992). Novel in-frame two codon translational hop during synthesis

of bovine placental lactogen in a recombinant strain of Escherichia coli. Nucleic Acids Res 20(24), 6707-12.

223. Kang, M. S., Hung, S. C., and Kieff, E. (2001). Epstein-Barr virus nuclear antigen 1 activates transcription from episomal but not integrated DNA and does not alter lymphocyte growth. Proc Natl Acad Sci U S A 98(26), 15233-8.

224. Kang, M. S., Lu, H. X., Yasui, T., Sharpe, A., Warren, H., Cahir-McFarland, E., Bronson, R., Hung, S. C., and Kieff, E. (2005). Epstein-Barr virus nuclear antigen 1 does not induce lymphoma in transgenic FVB mice. Proceedings of the National Academy of Sciences of the United States of America 102(3), 820-825.

225. Kang, M. S., Soni, V., Bronson, R., and Kieff, E. (2008). Epstein-Barr virus nuclear antigen 1 does not cause lymphoma in C57BL/6J mice. Journal of Virology 82(8), 4180-4183.

226. Kapoor, P., and Frappier, L. (2003). EBNA1 partitions Epstein-Barr virus plasmids in yeast cells by attaching to human EBNA1-binding protein 2 on mitotic chromosomes. J Virol 77(12), 6946-56.

227. Kapoor, P., Lavoie, B. D., and Frappier, L. (2005). EBP2 plays a key role in Epstein-Barr virus mitotic segregation and is regulated by aurora family kinases. Mol Cell Biol 25(12), 4934-45.

228. Kapoor, P., Shire, K., and Frappier, L. (2001). Reconstitution of Epstein-Barr virus-based plasmid partitioning in budding yeast. EMBO J 20(1-2), 222-30.

229. Karran, L., Gao, Y., Smith, P. R., and Griffin, B. E. (1992). Expression of a family of complementary-strand transcripts in Epstein-Barr virus-infected cells. Proc Natl Acad Sci U S A 89(17), 8058-62.

230. Kasprzewska, A. (2003). Plant chitinases--regulation and function. Cell Mol Biol Lett 8(3), 809-24.

231. Kaul, R., Murakami, M., Choudhuri, T., and Robertson, E. S. (2007). Epstein-Barr virus latent nuclear antigens can induce metastasis in a nude mouse model. Journal of Virology 81(19), 10352-10361.

232. Kaye, K. M., Izumi, K. M., and Kieff, E. (1993). Epstein-Barr virus latent membrane protein 1 is essential for B-lymphocyte growth transformation. Proc Natl Acad Sci U S A 90(19), 9150-4.

233. Kennedy, G., Komano, J., and Sugden, B. (2003). Epstein-Barr virus provides a survival factor to Burkitt's lymphomas. Proceedings of the National Academy of Sciences of the United States of America 100(24), 14269-14274.

234. Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L., and Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 37(Database issue), D387-92.

235. Kieff, E and Rickinson, AB: Epstein-Barr Virus and its replication. In Fields Virology. 5th edition. Edited by Fields BN, Knipe DM, Howley PM. Lippincott-Williams & Wilkins Publishers: Philadelphia; 2007:2603-2654.

236. Kim, S., and Lee, S. B. (2006). Rare codon clusters at 5'-end influence heterologous expression of archaeal gene in Escherichia coli. Protein Expr Purif 50(1), 49-57.

237. Kimura, H., Hoshino, Y., Kanegane, H., Tsuge, I., Okamura, T., Kawa, K., and Morishima, T. (2001). Clinical and virologic characteristics of chronic active Epstein-Barr virus infection. Blood 98(2), 280-6.

238. Kirchmaier, A. L., and Sugden, B. (1995). Plasmid maintenance of derivatives of oriP of Epstein-Barr virus. J Virol 69(2), 1280-3.

239. Knight, J. S., Lan, K., Bajaj, B., Sharma, N., Tsai, D. E., and Robertson, E. S. (2006). A peptide-based inhibitor for prevention of B cell hyperproliferation induced by Epstein-Barr virus. Virology 354(1), 207-14.

240. Komander, D. (2010). Mechanism, specificity and structure of the deubiquitinases. Subcell Biochem 54, 69-87.

241. Komander, D., Clague, M. J., and Urbe, S. (2009). Breaking the chains: structure and function of the deubiquitinases. Nat Rev Mol Cell Biol 10(8), 550-63.

242. Koonin, E. V., Wolf, Y. I., and Karev, G. P. (2002). The structure of the protein universe and genome evolution. Nature 420(6912), 218-23.

243. Kozakov, D., Hall, D. R., Beglov, D., Brenke, R., Comeau, S. R., Shen, Y., Li, K., Zheng, J., Vakili, P., Paschalidis, I., and Vajda, S. (2010). Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19. Proteins 78(15), 3124-30.

244. Kramer, A., and Schneider-Mergener, J. (1998). Synthesis and screening of peptide libraries on continuous cellulose membrane supports. Methods Mol Biol 87, 25-39.

245. Kubota, T., Miyamoto, K., Yasuda, M., Inamori, Y., and Tsujibo, H. (2004). Molecular characterization of an intracellular beta-N-acetylglucosaminidase involved in the chitin degradation system of Streptomyces thermoviolaceus OPC-520. Biosci Biotechnol Biochem 68(6), 1306-14.

246. Kumar, J., Ethayathulla, A. S., Srivastava, D. B., Singh, N., Sharma, S., Kaur, P., Srinivasan, A., and Singh, T. P. (2007). Carbohydrate-binding properties of goat secretory glycoprotein (SPG-40) and its functional implications: structures of the native glycoprotein and its four complexes with chitin-like oligosaccharides. Acta Crystallogr D Biol Crystallogr 63(Pt 4), 437-46.

247. Kung, C. P., Meckes, D. G., Jr., and Raab-Traub, N. (2011). Epstein-Barr virus LMP1 activates EGFR, STAT3, and ERK through effects on PKCdelta. J Virol 85(9), 4399-408.

248. Kuraku, S., Meyer, A., and Kuratani, S. (2009). Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? Mol Biol Evol 26(1), 47-59.

249. Kusserow, A., Pang, K., Sturm, C., Hrouda, M., Lentfer, J., Schmidt, H. A., Technau, U., von Haeseler, A., Hobmayer, B., Martindale, M. Q., and Holstein, T. W. (2005). Unexpected complexity of the Wnt gene family in a sea anemone. Nature 433(7022), 156-60.

250. Kwun, H. J., da Silva, S. R., Qin, H., Ferris, R. L., Tan, R., Chang, Y., and Moore, P. S. (2011). The central repeat domain 1 of Kaposi's sarcoma-associated herpesvirus (KSHV) latency associated-nuclear antigen 1 (LANA1) prevents cis MHC class I peptide presentation. Virology 412(2), 357-65.

251. Kzhyshkowska, J., Mamidi, S., Gratchev, A., Kremmer, E., Schmuttermaier, C., Krusell, L., Haus, G., Utikal, J., Schledzewski, K., Scholtze, J., and Goerdt, S. (2006). Novel stabilin-1 interacting chitinase-like protein (SI-CLP) is up-regulated in alternatively activated macrophages and secreted via lysosomal pathway. Blood 107(8), 3221-8.

252. Labute, P. (2008). The generalized Born/volume integral implicit solvent model: estimation of the free energy of hydration using London dispersion instead of atomic surface area. J Comput Chem 29, 1693-1698.

253. Lahn, B. T., and Page, D. C. (1999). Four evolutionary strata on the human X chromosome. Science 286(5441), 964-7.

254. Lee, C. G. (2009). Chitin, chitinases and chitinase-like proteins in allergic inflammation and tissue remodeling. Yonsei Med J 50(1), 22-30.

255. Lee, C. G., Da Silva, C. A., Dela Cruz, C. S., Ahangari, F., Ma, B., Kang, M. J., He, C. H., Takyar, S., and Elias, J. A. (2011). Role of chitin and chitinase/chitinase-like proteins in inflammation, tissue remodeling, and injury. Annu Rev Physiol 73, 479-501.

256. Lee, J. Y., and Spicer, A. P. (2000). Hyaluronan: a multifunctional, megaDalton, stealth molecule. Curr Opin Cell Biol 12(5), 581-6.

257. Lee, K. Y., Fu, H., Aladjem, M. I., and Myung, K. (2013). ATAD5 regulates the lifespan of DNA replication factories by modulating PCNA level on the chromatin. Journal of Cell Biology 200(1), 31-44.

258. Leifert, J. A., Holler, P. D., Harkins, S., Kranz, D. M., and Whitton, J. L. (2003). The cationic region from HIV tat enhances the cell-surface expression of epitope/MHC class I complexes. Gene Therapy 10(25), 2067-73.

259. Lerner, M. R., Andrews, N. C., Miller, G., and Steitz, J. A. (1981). Two small RNAs encoded by Epstein-Barr virus and complexed with protein are precipitated by antibodies from patients with systemic lupus erythematosus. Proc Natl Acad Sci U S A 78(2), 805-9.

260. Levin, L. I., Munger, K. L., O'Reilly, E. J., Falk, K. I., and Ascherio, A. (2010). Primary infection with the Epstein-Barr virus and risk of multiple sclerosis. Ann Neurol 67(6), 824-30.

261. Levitskaya, J., Sharipo, A., Leonchiks, A., Ciechanover, A., and Masucci, M. G. (1997). Inhibition of ubiquitin/proteasome-dependent protein degradation by the Gly-Ala repeat domain of the Epstein-Barr virus nuclear antigen 1. Proc Natl Acad Sci U S A 94(23), 12616-21.

262. Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. J Mol Biol 226(2), 507-33.

263. Li, H., and Greene, L. H. (2010). Sequence and structural analysis of the chitinase insertion domain reveals two conserved motifs involved in chitin-binding. PLoS One 5(1), e8654.

264. Li, H. P., and Chang, Y. S. (2003). Epstein-Barr virus latent membrane protein 1: structure and functions. J Biomed Sci 10(5), 490-504.

265. Li, J., D'Angiolella, V., Seeley, E. S., Kim, S., Kobayashi, T., Fu, W., Campos, E. I., Pagano, M., and Dynlacht, B. D. (2013). USP33 regulates centrosome biogenesis via deubiquitination of the centriolar protein CP110. Nature 495(7440), 255-9.

266. Li, M., Brooks, C. L., Kon, N., and Gu, W. (2004). A dynamic role of HAUSP in the p53-Mdm2 pathway. Mol Cell 13(6), 879-86.

267. Li, N., Thompson, S., Schultz, D. C., Zhu, W., Jiang, H., Luo, C., and Lieberman, P. M. (2010). Discovery of selective inhibitors against EBNA1 via high throughput in silico virtual screening. PLoS One 5(4), e10126.

268. Li, W., Bengtson, M. H., Ulbrich, A., Matsuda, A., Reddy, V. A., Orth, A., Chanda, S. K., Batalov, S., and Joazeiro, C. A. (2008). Genome-wide and functional annotation of human E3 ubiquitin ligases identifies MULAN, a mitochondrial E3 that regulates the organelle's dynamics and signaling. PLoS One 3(1), e1487.

269. Li, Y., and Zhang, Y. (2009). REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins 76(3), 665-76.

270. Lin, A., Wang, S., Nguyen, T., Shire, K., and Frappier, L. (2008). The EBNA1 protein of Epstein-Barr virus functionally interacts with Brd4. J Virol 82(24), 12009-19.

271. Lindsay, L. L., Wieduwilt, M. J., and Hedrick, J. L. (1999). Oviductin, the Xenopus laevis oviductal protease that processes egg envelope glycoprotein gp43, increases sperm binding to envelopes, and is translated as part of an unusual mosaic protein composed of two protease and several CUB domains. Biol Reprod 60(4), 989-95.

272. Liu, C. D., Chen, Y. L., Min, Y. L., Zhao, B., Cheng, C. P., Kang, M. S., Chiu, S. J., Kieff, E., and Peng, C. W. (2012). The nuclear chaperone nucleophosmin escorts an Epstein-Barr Virus nuclear antigen to establish transcriptional cascades for latent infection in human B cells. PLoS Pathog 8(12), e1003084.

273. Liu, J., Faeder, J. R., and Camacho, C. J. (2009). Toward a quantitative theory of intrinsically disordered proteins and their function. Proc Natl Acad Sci U S A 106(47), 19819-23.

274. Liu, X., Fan, K., and Wang, W. (2004). The number of protein folds and their distribution over families in nature. Proteins 54(3), 491-9.

275. Long, H., Sabatier, C., Ma, L., Plump, A., Yuan, W., Ornitz, D. M., Tamada, A., Murakami, F., Goodman, C. S., and Tessier-Lavigne, M. (2004). Conserved roles for Slit and Robo proteins in midline commissural axon guidance. Neuron 42(2), 213-23.

276. Longnecker, R. (2000). Epstein-Barr virus latency: LMP2, a regulator or means for Epstein-Barr virus persistence? Adv Cancer Res 79, 175-200.

277. Longnecker, R., and Kieff, E. (1990). A second Epstein-Barr virus membrane protein (LMP2) is expressed in latent infection and colocalizes with LMP1. J Virol 64(5), 2319-26.

278. Longnecker, R., Miller, C. L., Miao, X. Q., Tomkinson, B., and Kieff, E. (1993). The last seven transmembrane and carboxy-terminal cytoplasmic domains of Epstein-Barr virus latent membrane protein 2 (LMP2) are dispensable for lymphocyte infection and growth transformation in vitro. J Virol 67(4), 2006-13.

279. Longnecker, R., Miller, C. L., Tomkinson, B., Miao, X. Q., and Kieff, E. (1993). Deletion of DNA encoding the first five transmembrane domains of Epstein-Barr virus latent membrane proteins 2A and 2B. J Virol 67(8), 5068-74.

280. Loytynoja, A., Vilella, A. J., and Goldman, N. (2012). Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics 28(13), 1684-91.

281. Lu, F., Wikramasinghe, P., Norseen, J., Tsai, K., Wang, P., Showe, L., Davuluri, R. V., and Lieberman, P. M. (2010). Genome-wide analysis of host-chromosome binding sites for Epstein-Barr Virus Nuclear Antigen 1 (EBNA1). Virol J 7, 262.

282. Lu, J., Murakami, M., Verma, S. C., Cai, Q., Haldar, S., Kaul, R., Wasik, M. A., Middeldorp, J., and Robertson, E. S. (2011). Epstein-Barr Virus nuclear antigen 1 (EBNA1) confers resistance to apoptosis in EBV-positive B-lymphoma cells through up-regulation of survivin. Virology 410(1), 64-75.

283. Lui, T. T., Lacroix, C., Ahmed, S. M., Goldenberg, S. J., Leach, C. A., Daulat, A. M., and Angers, S. (2011). The ubiquitin-specific protease USP34 regulates axin stability and Wnt/beta-catenin signaling. Mol Cell Biol 31(10), 2053-65.

284. Luo, W., Li, Y., Tang, C. H., Abruzzi, K. C., Rodriguez, J., Pescatore, S., and Rosbash, M. (2012). CLOCK deubiquitylation by USP8 inhibits CLK/CYC transcription in Drosophila. Genes Dev 26(22), 2536-49.

285. Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. Science 290(5494), 1151-5.

286. Mackey, D., Middleton, T., and Sugden, B. (1995). Multiple regions within EBNA1 can link DNAs. J Virol 69(10), 6199-208.

287. Maeda, E., Akahane, M., Kiryu, S., Kato, N., Yoshikawa, T., Hayashi, N., Aoki, S., Minami, M., Uozaki, H., Fukayama, M., and Ohtomo, K. (2009). Spectrum of Epstein-Barr virus-related diseases: a pictorial review. Jpn J Radiol 27(1), 4-19.

288. Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011, bar009.

289. Maines-Bandiera, S., Woo, M. M., Borugian, M., Molday, L. L., Hii, T., Gilks, B., Leung, P. C., Molday, R. S., and Auersperg, N. (2010). Oviductal glycoprotein (OVGP1, MUC9): a differentiation-based mucin present in serum of women with ovarian cancer. Int J Gynecol Cancer 20(1), 16-22.

290. Malakhov, M. P., Malakhova, O. A., Kim, K. I., Ritchie, K. J., and Zhang, D. E. (2002). UBP43 (USP18) specifically removes ISG15 from conjugated proteins. J Biol Chem 277(12), 9976-81.

291. Malik-Soni, N., and Frappier, L. (2012). Proteomic profiling of EBNA1-host protein interactions in latent and lytic Epstein-Barr virus infections. J Virol 86(12), 6999-7002.

292. Malmstrom, L., Riffle, M., Strauss, C. E., Chivian, D., Davis, T. N., Bonneau, R., and Baker, D. (2007). Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. PLoS Biol 5(4), e76.

293. Marchler-Bauer, A., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Liebert, C. A., Liu, C., Lu, F., Lu, S., Marchler, G. H., Mullokandov, M., Song, J. S., Tasneem, A., Thanki, N., Yamashita, R. A., Zhang, D., Zhang, N., and Bryant, S. H. (2009). CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 37(Database issue), D205-10.

294. Marques-Bonet, T., Girirajan, S., and Eichler, E. E. (2009). The origins and impact of primate segmental duplications. Trends Genet 25(10), 443-54.

295. Martinez, E., Palhan, V. B., Tjernberg, A., Lymar, E. S., Gamper, A. M., Kundu, T. K., Chait, B. T., and Roeder, R. G. (2001). Human STAGA complex is a chromatin-acetylating transcription coactivator that interacts with pre-mRNA splicing and DNA damage-binding factors in vivo. Mol Cell Biol 21(20), 6782-95.

296. Martus, N. S., Verhage, H. G., Mavrogianis, P. A., and Thibodeaux, J. K. (1998). Enhancement of bovine oocyte fertilization in vitro with a bovine oviductal specific glycoprotein. J Reprod Fertil 113(2), 323-9.

297. Maruo, S., Johannsen, E., Illanes, D., Cooper, A., and Kieff, E. (2003). Epstein-Barr Virus nuclear protein EBNA3A is critical for maintaining lymphoblastoid cell line growth. J Virol 77(19), 10437-47.

298. Maruo, S., Johannsen, E., Illanes, D., Cooper, A., Zhao, B., and Kieff, E. (2005). Epstein-Barr virus nuclear protein 3A domains essential for growth of lymphoblasts: transcriptional regulation through RBP-Jkappa/CBF1 is critical. J Virol 79(16), 10171-9.

299. Massoumi, R., and Paus, R. (2007). Cylindromatosis and the CYLD gene: new lessons on the molecular principles of epithelial growth control. Bioessays 29(12), 1203-14.

300. McCullough, J., Clague, M. J., and Urbe, S. (2004). AMSH is an endosome-associated ubiquitin isopeptidase. J Cell Biol 166(4), 487-92.

301. McGeoch, D. J., Rixon, F. J., and Davison, A. J. (2006). Topics in herpesvirus genomics and evolution. Virus Res 117(1), 90-104.

302. Meierhofer, D., Wang, X., Huang, L., and Kaiser, P. (2008). Quantitative analysis of global ubiquitination in HeLa cells by mass spectrometry. J Proteome Res 7(10), 4566-76.

303. Meng, G., Zhao, Y., Bai, X., Liu, Y., Green, T. J., Luo, M., and Zheng, X. (2010). Structure of human stabilin-1 interacting chitinase-like protein (SI-CLP) reveals a saccharide-binding cleft with lower sugar-binding selectivity. J Biol Chem 285(51), 39898-904.

304. Meyer, A., and Van de Peer, Y. (2005). From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays 27(9), 937-45.

305. Meyer, M. F., and Kreil, G. (1996). Cells expressing the DG42 gene from early Xenopus embryos synthesize hyaluronan. Proc Natl Acad Sci U S A 93(10), 4543-7.

306. Mohanty, A. K., Singh, G., Paramasivam, M., Saravanan, K., Jabeen, T., Sharma, S., Yadav, S., Kaur, P., Kumar, P., Srinivasan, A., and Singh, T. P. (2003). Crystal structure of a novel regulatory 40-kDa mammary gland protein (MGP-40) secreted during involution. J Biol Chem 278(16), 14451-60.

307. Molesworth, S. J., Lake, C. M., Borza, C. M., Turk, S. M., and Hutt-Fletcher, L. M. (2000). Epstein-Barr virus gH is essential for penetration of B cells but also plays a role in attachment of virus to epithelial cells. J Virol 74(14), 6324-32.

308. Moody, C. A., Scott, R. S., Su, T., and Sixbey, J. W. (2003). Length of Epstein-Barr virus termini as a determinant of epithelial cell clonal emergence. J Virol 77(15), 8555-61.

309. Moriyama, K., Yoshizawa-Sugata, N., Obuse, C., Tsurimoto, T., and Masai, H. (2012). Epstein-Barr nuclear antigen 1 (EBNA1)-dependent recruitment of origin recognition complex (Orc) on oriP of Epstein-Barr virus with purified proteins: stimulation by Cdc6 through its direct interaction with EBNA1. J Biol Chem 287(28), 23977-94.

310. Murai, J., Yang, K., Dejsuphong, D., Hirota, K., Takeda, S., and D'Andrea, A. D. (2011). The USP1/UAF1 complex promotes double-strand break repair through homologous recombination. Mol Cell Biol 31(12), 2462-9.

311. Murakami, M., Lan, K., Subramanian, C., and Robertson, E. S. (2005). Epstein-Barr virus nuclear antigen 1 interacts with Nm23-H1 in lymphoblastoid cell lines and inhibits its ability to suppress cell migration. J Virol 79(3), 1559-68.

312. Murray, R. Z., Jolly, L. A., and Wood, S. A. (2004). The FAM deubiquitylating enzyme localizes to multiple points of protein trafficking in epithelia, where it associates with E-cadherin and beta-catenin. Mol Biol Cell 15(4), 1591-9.

313. Nacher, J. C., Hayashida, M., and Akutsu, T. (2010). The role of internal duplication in the evolution of multi-domain proteins. Biosystems 101(2), 127-35.

314. Nagano, N., Orengo, C. A., and Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 321(5), 741-65.

315. Nakagawa, T., Kajitani, T., Togo, S., Masuko, N., Ohdan, H., Hishikawa, Y., Koji, T., Matsuyama, T., Ikura, T., Muramatsu, M., and Ito, T. (2008). Deubiquitylation of histone H2A activates transcriptional initiation via trans-histone cross-talk with H3K4 di- and trimethylation. Genes Dev 22(1), 37-49.

316. Nakatani, Y., Takeda, H., Kohara, Y., and Morishita, S. (2007). Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res 17(9), 1254-65.

317. Nakayama, T., Fujisawa, R., Izawa, D., Hieshima, K., Takada, K., and Yoshie, O. (2002). Human B cells immortalized with Epstein-Barr virus upregulate CCR6 and CCR10 and downregulate CXCR4 and CXCR5. J Virol 76(6), 3072-7.

318. Nanbo, A., and Takada, K. (2002). The role of Epstein-Barr virus-encoded small RNAs (EBERs) in oncogenesis. Rev Med Virol 12(5), 321-6.

319. Nanbo, A., Yoshiyama, H., and Takada, K. (2005). Epstein-Barr virus-encoded poly(A)- RNA confers resistance to apoptosis mediated through Fas by blocking the PKR pathway in human epithelial intestine 407 cells. J Virol 79(19), 12280-5.

320. Nayyar, V. K., Shire, K., and Frappier, L. (2009). Mitotic chromosome interactions of Epstein-Barr nuclear antigen 1 (EBNA1) and human EBNA1-binding protein 2 (EBP2). J Cell Sci 122(Pt 23), 4341-50.

321. Nei, M., and Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39, 121-52.

322. Newell, G. R., and Stevens, D. A. (1971). Epstein-Barr virus antibody in systemic lupus erythematosus. Lancet 1(7700), 652.

323. Ney, J. T., Zhou, H., Sipos, B., Buttner, R., Chen, X., Kloppel, G., and Gutgemann, I. (2007). Podocalyxin-like protein 1 expression is useful to differentiate pancreatic ductal adenocarcinomas from adenocarcinomas of the biliary and gastrointestinal tracts. Hum Pathol 38(2), 359-64.

324. Nicassio, F., Corrado, N., Vissers, J. H., Areces, L. B., Bergink, S., Marteijn, J. A., Geverts, B., Houtsmuller, A. B., Vermeulen, W., Di Fiore, P. P., and Citterio, E. (2007). Human USP3 is a chromatin modifier required for S phase progression and genome stability. Curr Biol 17(22), 1972-7.

325. Nichols, R. (2001). Gene trees and species trees are not the same. Trends Ecol Evol 16(7), 358-364.

326. Nielsen, J. S., and McNagny, K. M. (2008). Novel functions of the CD34 family. J Cell Sci 121(Pt 22), 3683-92.

327. Nielsen, J. S., and McNagny, K. M. (2008). Novel functions of the CD34 family. J Cell Sci 121(Pt 22), 3683-92.

328. Nijman, S. M., Huang, T. T., Dirac, A. M., Brummelkamp, T. R., Kerkhoven, R. M., D'Andrea, A. D., and Bernards, R. (2005a). The deubiquitinating enzyme USP1 regulates the Fanconi anemia pathway. Mol Cell 17(3), 331-9.

329. Nijman, S. M., Luna-Vargas, M. P., Velds, A., Brummelkamp, T. R., Dirac, A. M., Sixma, T. K., and Bernards, R. (2005b). A genomic and functional inventory of deubiquitinating enzymes. Cell 123(5), 773-86.

330. Nikoskelainen, J., Panelius, M., and Salmi, A. (1972). E.B. virus and multiple sclerosis. Br Med J 4(5832), 111.

331. Niller HH, M. J. (2012). Similarities between the epstein-barr virus (ebv) nuclear protein ebna1 and the pioneer transcription factor foxa: Is ebna1 a “bookmarking” oncoprotein that alters the host cell epigenotype. Pathogens 1, 15.

332. Nio, J., Fujimoto, W., Konno, A., Kon, Y., Owhashi, M., and Iwanaga, T. (2004). Cellular expression of murine Ym1 and Ym2, chitinase family proteins, as revealed by in situ hybridization and immunohistochemistry. Histochem Cell Biol 121(6), 473-82.

333. Nitsche, F., Bell, A., and Rickinson, A. (1997). Epstein-Barr virus leader protein enhances EBNA-2-mediated transactivation of latent membrane protein 1 expression: a role for the W1W2 repeat domain. J Virol 71(9), 6619-28.

334. Noguchi, T., Ishii, K., Fukutomi, H., Naguro, I., Matsuzawa, A., Takeda, K., and Ichijo, H. (2008). Requirement of reactive oxygen species-dependent activation of ASK1-p38 MAPK pathway for extracellular ATP-induced apoptosis in macrophage. J Biol Chem 283(12), 7657-65.

335. Norseen, J., Johnson, F. B., and Lieberman, P. M. (2009). Role for G-quadruplex RNA binding by Epstein-Barr virus nuclear antigen 1 in DNA replication and metaphase chromosome attachment. J Virol 83(20), 10336-46.

336. Norseen, J., Thomae, A., Sridharan, V., Aiyar, A., Schepers, A., and Lieberman, P. M. (2008). RNA-dependent recruitment of the origin recognition complex. EMBO J 27(22), 3024-35.

337. Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302(1), 205-17.

338. Ober, C., and Chupp, G. L. (2009). The chitinase and chitinase-like proteins: a review of genetic and functional studies in asthma and immune-mediated diseases. Curr Opin Allergy Clin Immunol 9(5), 401-8.

339. Odumade, O. A., Hogquist, K. A., and Balfour, H. H., Jr. (2011). Progress and problems in understanding and managing primary Epstein-Barr virus infections. Clin Microbiol Rev 24(1), 193-209.

340. Ohga, S., Nomura, A., Takada, H., and Hara, T. (2002). Immunological aspects of Epstein-Barr virus infection. Crit Rev Oncol Hematol 44(3), 203-15.

341. Ohno, T., Armand, S., Hata, T., Nikaidou, N., Henrissat, B., Mitsutomi, M., and Watanabe, T. (1996). A modular family 19 chitinase found in the prokaryotic organism Streptomyces griseus HUT 6037. J Bacteriol 178(17), 5065-70.

342. Ojima, H., Fukuda, T., Nakajima, T., Takenoshita, S., and Nagamachi, Y. (1996). Discrepancy between clinical and pathological lymph node evaluation in Epstein-Barr virus-associated gastric cancers. Anticancer Res 16(5B), 3081-4.

343. Okada, H., Uezu, A., Soderblom, E. J., Moseley, M. A., 3rd, Gertler, F. B., and Soderling, S. H. (2012). Peptide array X-linking (PAX): a new peptide-protein identification approach. PLoS One 7(5), e37035.

344. Oliveira, A. M., Chou, M. M., Perez-Atayde, A. R., and Rosenberg, A. E. (2006). Aneurysmal bone cyst: a neoplasm driven by upregulation of the USP6 oncogene. J Clin Oncol 24(1), e1; author reply e2.

345. Olland, A. M., Strand, J., Presman, E., Czerwinski, R., Joseph-McCarthy, D., Krykbaev, R., Schlingmann, G., Chopra, R., Lin, L., Fleming, M., Kriz, R., Stahl, M., Somers, W., Fitz, L., and Mosyak, L. (2009). Triad of polar residues implicated in pH specificity of acidic mammalian chitinase. Protein Sci 18(3), 569-78.

346. O'Neil, J. D., Owen, T. J., Wood, V. H., Date, K. L., Valentine, R., Chukwuma, M. B., Arrand, J. R., Dawson, C. W., and Young, L. S. (2008). Epstein-Barr virus-encoded EBNA1 modulates the AP-1 transcription factor pathway in nasopharyngeal carcinoma cells and enhances angiogenesis in vitro. J Gen Virol 89(Pt 11), 2833-42.

347. O'Nions, J., Turner, A., Craig, R., and Allday, M. J. (2006). Epstein-Barr virus selectively deregulates DNA damage responses in normal B cells but has no detectable effect on regulation of the tumor suppressor p53. Journal of Virology 80(24), 12408-12413.

348. Orlowski, R., Polvino-Bodnar, M., Hearing, J., and Miller, G. (1990). Inhibition of specific binding of EBNA 1 to DNA by murine monoclonal and certain human polyclonal antibodies. Virology 176(2), 638–642.

349. Otto, S. P. (2007). The evolutionary consequences of polyploidy. Cell 131(3), 452-62.

350. Page, R. D. (2013). BioNames: linking taxonomy, texts, and trees. PeerJ 1, e190.

351. Peng, J., Schwartz, D., Elias, J. E., Thoreen, C. C., Cheng, D., Marsischky, G., Roelofs, J., Finley, D., and Gygi, S. P. (2003). A proteomics approach to understanding protein ubiquitination. Nat Biotechnol 21(8), 921-6.

352. Perelman, P., Johnson, W. E., Roos, C., Seuanez, H. N., Horvath, J. E., Moreira, M. A., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M. P., Silva, A., O'Brien, S. J., and Pecon-Slattery, J. (2011). A molecular phylogeny of living primates. PLoS Genet 7(3), e1001342.

353. Peretti, T., Waisberg, J., Mader, A. M., de Matos, L. L., da Costa, R. B., Conceicao, G. M., Lopes, A. C., Nader, H. B., and Pinhal, M. A. (2008). Heparanase-2, syndecan-1, and extracellular matrix remodeling in colorectal carcinoma. Eur J Gastroenterol Hepatol 20(8), 756-65.

354. Pfaff, J., Hennig, J., Herzog, F., Aebersold, R., Sattler, M., Niessing, D., and Meister, G. (2013). Structural features of Argonaute-GW182 protein interactions. Proc Natl Acad Sci U S A 110(40), E3770-E3779.

355. Pickart, C. M. (2001). Mechanisms underlying ubiquitination. Annu Rev Biochem 70, 503-33.

356. Pinhal, D., Yoshimura, T. S., Araki, C. S., and Martins, C. (2011). The 5S rDNA family evolves through concerted and birth-and-death evolution in fish genomes: an example from freshwater stingrays. BMC Evol Biol 11, 151.

357. Plazinski, W., and Knys-Dzieciuch, A. (2012). Interactions between CD44 protein and hyaluronan: insights from the computational study. Mol Biosyst 8(2), 543-7.

358. Poblete Gutierrez, P., Eggermann, T., Holler, D., Jugert, F. K., Beermann, T., Grussendorf-Conen, E. I., Zerres, K., Merk, H. F., and Frank, J. (2002). Phenotype diversity in familial cylindromatosis: a frameshift mutation in the tumor suppressor gene CYLD underlies different tumors of skin appendages. J Invest Dermatol 119(2), 527-31.

359. Postlethwait, J. H., Yan, Y. L., Gates, M. A., Horne, S., Amores, A., Brownlie, A., Donovan, A., Egan, E. S., Force, A., Gong, Z., Goutel, C., Fritz, A., Kelsh, R., Knapik, E., Liao, E., Paw, B., Ransom, D., Singer, A., Thomson, M., Abduljabbar, T. S., Yelick, P., Beier, D., Joly, J. S., Larhammar, D., Rosa, F., Westerfield, M.,

Zon, L. I., Johnson, S. L., and Talbot, W. S. (1998). Vertebrate genome evolution and the zebrafish gene map. Nat Genet 18(4), 345-9.

360. Potu, H., Sgorbissa, A., and Brancolini, C. (2010). Identification of USP18 as an important regulator of the susceptibility to IFN-alpha and drug-induced apoptosis. Cancer Res 70(2), 655-65.

361. Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., Silman, I., and Sussman, J. L. (2005). FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21(16), 3435-8.

362. Prota, A. E., Sage, D. R., Stehle, T., and Fingeroth, J. D. (2002). The crystal structure of human CD21: Implications for Epstein-Barr virus and C3d binding. Proc Natl Acad Sci U S A 99(16), 10641-6.

363. Putnam, N. H., Butts, T., Ferrier, D. E., Furlong, R. F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J. K., Benito-Gutierrez, E. L., Dubchak, I., Garcia-Fernandez, J., Gibson-Brown, J. J., Grigoriev, I. V., Horton, A. C., de Jong, P. J., Jurka, J., Kapitonov, V. V., Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio, L. A., Salamov, A. A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin, I. T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L. Z., Holland, P. W., Satoh, N., and Rokhsar, D. S. (2008). The amphioxus genome and the evolution of the chordate karyotype. Nature 453(7198), 1064-71.

364. Qiu, X. B., Markant, S. L., Yuan, J., and Goldberg, A. L. (2004). Nrdp1-mediated degradation of the gigantic IAP, BRUCE, is a novel pathway for triggering apoptosis. EMBO J 23(4), 800-10.

365. Qureshi, A. M., Hannigan, A., Campbell, D., Nixon, C., and Wilson, J. B. (2011). Chitinase-like proteins are autoantigens in a model of inflammation-promoted incipient neoplasia. Genes Cancer 2(1), 74-87.

366. Raab-Traub, N., and Flynn, K. (1986). The structure of the termini of the Epstein-Barr virus as a marker of clonal cellular proliferation. Cell 47(6), 883-9.

367. Ramakrishna, S., Suresh, B., and Baek, K. H. (2011). The role of deubiquitinating enzymes in apoptosis. Cell Mol Life Sci 68(1), 15-26.

368. Rawlins, D. R., Milman, G., Hayward, S. D., and Hayward, G. S. (1985). Sequence-specific DNA binding of the Epstein-Barr virus nuclear antigen (EBNA-1) to clustered sites in the plasmid maintenance region. Cell 42(3), 859-68.

369. Reese, T. A., Liang, H. E., Tager, A. M., Luster, A. D., Van Rooijen, N., Voehringer, D., and Locksley, R. M. (2007). Chitin induces accumulation in tissue of innate immune cells associated with allergy. Nature 447(7140), 92-6.

370. Reichart, P. A., Langford, A., Gelderblom, H. R., Pohle, H. D., Becker, J., and Wolf, H. (1989). Oral hairy leukoplakia: observations in 95 cases and review of the literature. J Oral Pathol Med 18(7), 410-5.

371. Reisman, D., Yates, J., and Sugden, B. (1985). A putative origin of replication of plasmids derived from Epstein-Barr virus is composed of two cis-acting components. Mol Cell Biol 5(8), 1822-32.

372. Reissig, S., Hovelmeyer, N., Weigmann, B., Nikolaev, A., Kalt, B., Wunderlich, T. F., Hahn, M., Neurath, M. F., and Waisman, A. (2012). The tumor suppressor CYLD controls the function of murine regulatory T cells. J Immunol 189(10), 4770-6.

373. Repellin, C. E., Tsimbouri, P. M., Philbey, A. W., and Wilson, J. B. (2010). Lymphoid hyperplasia and lymphoma in transgenic mice expressing the small non-coding RNA, EBER1 of Epstein-Barr virus. PLoS One 5(2), e9092.

374. Reyes-Turcu, F. E., Shanks, J. R., Komander, D., and Wilkinson, K. D. (2008). Recognition of polyubiquitin isoforms by the multiple ubiquitin binding modules of isopeptidase T. J Biol Chem 283(28), 19581-92.

375. Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 6(1), 1-6.

376. Rios, Y., Melmed, S., Lin, S., and Liu, N. A. (2011). Zebrafish usp39 mutation leads to rb1 mRNA splicing defect and pituitary lineage expansion. PLoS Genet 7(1), e1001271.

377. Ritchie, D. W. (2008). Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci 9(1), 1-15.

378. Rivailler, P., Cho, Y. G., and Wang, F. (2002). Complete genomic sequence of an Epstein-Barr virus-related herpesvirus naturally infecting a new world primate: a defining point in the evolution of oncogenic lymphocryptoviruses. J Virol 76(23), 12055-68.

379. Rojas, A. M., Fuentes, G., Rausell, A., and Valencia, A. (2012). The Ras protein superfamily: evolutionary tree and role of conserved amino acids. Journal of Cell Biology 196(2), 189-201.

380. Row, P. E., Prior, I. A., McCullough, J., Clague, M. J., and Urbe, S. (2006). The ubiquitin isopeptidase UBPY regulates endosomal ubiquitin dynamics and is essential for receptor down-regulation. J Biol Chem 281(18), 12618-24.

381. Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5(4), 725-38.

382. Roy, S. W. (2009). Phylogenomics: gene duplication, unrecognized paralogy and outgroup choice. PLoS One 4(2), e4568.

383. Ruf, I. K., Rhyne, P. W., Yang, C., Cleveland, J. L., and Sample, J. T. (2000). Epstein-Barr virus small RNAs potentiate tumorigenicity of Burkitt lymphoma cells independently of an effect on apoptosis. J Virol 74(21), 10223-8.

384. Ryan, J. L., Fan, H., Glaser, S. L., Schichman, S. A., Raab-Traub, N., and Gulley, M. L. (2004). Epstein-Barr virus quantitation by real-time PCR targeting multiple gene segments: a novel approach to screen for the virus in paraffin-embedded tissue and plasma. J Mol Diagn 6(4), 378-85.

385. Sali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3), 779-815.

386. Salomoni, P., Dvorkina, M., and Michod, D. (2012). Role of the promyelocytic leukaemia protein in cell death regulation. Cell Death Dis 3, e247.

387. Sample, J., Henson, E. B., and Sample, C. (1992). The Epstein-Barr virus nuclear protein 1 promoter active in type I latency is autoregulated. J Virol 66(8), 4654-61.

388. Sample, J., Hummel, M., Braun, D., Birkenbach, M., and Kieff, E. (1986). Nucleotide sequences of mRNAs encoding Epstein-Barr virus nuclear proteins: a probable transcriptional initiation site. Proc Natl Acad Sci U S A 83(14), 5096-100.

389. Sample, J., Liebowitz, D., and Kieff, E. (1989). Two related Epstein-Barr virus membrane proteins are encoded by separate genes. J Virol 63(2), 933-7.

390. Sample, J., Young, L., Martin, B., Chatman, T., Kieff, E., and Rickinson, A. (1990). Epstein-Barr virus types 1 and 2 differ in their EBNA-3A, EBNA-3B, and EBNA-3C genes. J Virol 64(9), 4084-92.

391. Sarasin, A. (2012). UVSSA and USP7: new players regulating transcription-coupled nucleotide excision repair in human cells. Genome Med 4(5), 44.

392. Saridakis, V., Sheng, Y., Sarkari, F., Holowaty, M. N., Shire, K., Nguyen, T., Zhang, R. G., Liao, J., Lee, W., Edwards, A. M., Arrowsmith, C. H., and Frappier, L. (2005). Structure of the p53 binding domain of HAUSP/USP7 bound to Epstein-Barr nuclear antigen 1 implications for EBV-mediated immortalization. Mol Cell 18(1), 25-36.

393. Sarkari, F., Sanchez-Alcaraz, T., Wang, S., Holowaty, M. N., Sheng, Y., and Frappier, L. (2009). EBNA1-mediated recruitment of a histone H2B

deubiquitylating complex to the Epstein-Barr virus latent origin of DNA replication. PLoS Pathog 5(10), e1000624.

394. Sarkari, F., Sheng, Y., and Frappier, L. (2010). USP7/HAUSP promotes the sequence-specific DNA binding activity of p53. PLoS One 5(9), e13040.

395. Sarkari, F., Wang, X., Nguyen, T., and Frappier, L. (2011). The herpesvirus associated ubiquitin specific protease, USP7, is a negative regulator of PML proteins and PML nuclear bodies. PLoS One 6(1), e16598.

396. Satoh, T., Abe, H., Sendai, Y., Iwata, H., and Hoshi, H. (1995). Biochemical characterization of a bovine oviduct-specific sialo-glycoprotein that sustains sperm viability in vitro. Biochim Biophys Acta 1266(2), 117-23.

397. Scaglioni, P. P., Yung, T. M., Cai, L. F., Erdjument-Bromage, H., Kaufman, A. J., Singh, B., Teruya-Feldstein, J., Tempst, P., and Pandolfi, P. P. (2006). A CK2-dependent mechanism for degradation of the PML tumor suppressor. Cell 126(2), 269-83.

398. Scaglioni, P. P., Yung, T. M., Choi, S., Baldini, C., Konstantinidou, G., and Pandolfi, P. P. (2008). CK2 mediates phosphorylation and ubiquitin-mediated degradation of the PML tumor suppressor. Mol Cell Biochem 316(1-2), 149-54.

399. Scheeff, E. D., and Bourne, P. E. (2005). Structural evolution of the protein kinase-like superfamily. PLoS Comput Biol 1(5), e49.

400. Schimpl, M., Rush, C. L., Betou, M., Eggleston, I. M., Recklies, A. D., and van Aalten, D. M. (2012). Human YKL-39 is a pseudo-chitinase with retained chitooligosaccharide-binding properties. Biochem J 446(1), 149-57.

401. Schlessinger, J., Plotnikov, A. N., Ibrahimi, O. A., Eliseenkova, A. V., Yeh, B. K., Yayon, A., Linhardt, R. J., and Mohammadi, M. (2000). Crystal structure of a ternary FGF-FGFR-heparin complex reveals a dual role for heparin in FGFR binding and dimerization. Mol Cell 6(3), 743-50.

402. Schmitz, R., Young, R. M., Ceribelli, M., Jhavar, S., Xiao, W. M., Zhang, M. Z., Wright, G., Shaffer, A. L., Hodson, D. J., Buras, E., Liu, X. L., Powell, J., Yang, Y. D., Xu, W. H., Zhao, H., Kohlhammer, H., Rosenwald, A., Kluin, P., Muller-Hermelink, H. K., Ott, G., Gascoyne, R. D., Connors, J. M., Rimsza, L. M., Campo, E., Jaffe, E. S., Delabie, J., Smeland, E. B., Ogwang, M. D., Reynolds, S. J., Fisher, R. I., Braziel, R. M., Tubbs, R. R., Cook, J. R., Weisenburger, D. D., Chan, W. C., Pittaluga, S., Wilson, W., Waldmann, T. A., Rowe, M., Mbulaiteye, S. M., Rickinson, A. B., and Staudt, L. M. (2012). Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. Nature 490(7418), 116-120.

403. Schmitz-Esser, S., Tischler, P., Arnold, R., Montanaro, J., Wagner, M., Rattei, T., and Horn, M. (2010). The genome of the amoeba symbiont "Candidatus Amoebophilus asiaticus" reveals common mechanisms for host cell interaction among amoeba-associated bacteria. J Bacteriol 192(4), 1045-57.

404. Schmitz-Esser, S., Toenshoff, E. R., Haider, S., Heinz, E., Hoenninger, V. M., Wagner, M., and Horn, M. (2008). Diversity of bacterial endosymbionts of environmental acanthamoeba isolates. Appl Environ Microbiol 74(18), 5822-31.

405. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005a). Geometry-based flexible and symmetric protein docking. Proteins 60(2), 224-31.

406. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005b). PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res 33(Web Server issue), W363-7.

407. Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5(12), e1000605.

408. Schwemmlein, M., Peipp, M., Barbin, K., Saul, D., Stockmeyer, B., Repp, R., Birkmann, J., Oduncu, F., Emmerich, B., and Fey, G. H. (2006). A CD33-specific

single-chain immunotoxin mediates potent apoptosis of cultured human myeloid leukaemia cells. Br J Haematol 133(2), 141-51.

409. Schwertman, P., Vermeulen, W., and Marteijn, J. A. (2013). UVSSA and USP7, a new couple in transcription-coupled DNA repair. Chromosoma 122(4), 275-84.

410. Sears, J., Ujihara, M., Wong, S., Ott, C., Middeldorp, J., and Aiyar, A. (2004). The amino terminus of Epstein-Barr Virus (EBV) nuclear antigen 1 contains AT hooks that facilitate the replication and partitioning of latent EBV genomes by tethering them to cellular chromosomes. J Virol 78(21), 11487-505.

411. Semino, C. E., Specht, C. A., Raimondi, A., and Robbins, P. W. (1996). Homologs of the Xenopus developmental gene DG42 are present in zebrafish and mouse and are involved in the synthesis of Nod-like chitin oligosaccharides during early embryogenesis. Proc Natl Acad Sci U S A 93(10), 4548-53.

412. Shao, R., Hamel, K., Petersen, L., Cao, Q. J., Arenas, R. B., Bigelow, C., Bentley, B., and Yan, W. (2009). YKL-40, a secreted glycoprotein, promotes tumor angiogenesis. Oncogene 28(50), 4456-68.

413. Sheng, W., Decaussin, G., Sumner, S., and Ooka, T. (2001). N-terminal domain of BARF1 gene encoded by Epstein-Barr virus is essential for malignant transformation of rodent fibroblasts and activation of BCL-2. Oncogene 20(10), 1176-85.

414. Sheu, L. F., Chen, A., Meng, C. L., Ho, K. C., Lee, W. H., Leu, F. J., and Chao, C. F. (1996). Enhanced malignant progression of nasopharyngeal carcinoma cells mediated by the expression of Epstein-Barr nuclear antigen 1 in vivo. Journal of Pathology 180(3), 243-248.

415. Shin, J. M., Yoo, K. J., Kim, M. S., Kim, D., and Baek, K. H. (2006). Hyaluronan- and RNA-binding deubiquitinating enzymes of USP17 family members associated with cell viability. BMC Genomics 7, 292.

416. Shire, K., Ceccarelli, D. F., Avolio-Hunter, T. M., and Frappier, L. (1999). EBP2, a human protein that interacts with sequences of the Epstein-Barr virus nuclear antigen 1 important for plasmid maintenance. J Virol 73(4), 2587-95.

417. Shire, K., Kapoor, P., Jiang, K., Hing, M. N., Sivachandran, N., Nguyen, T., and Frappier, L. (2006). Regulation of the EBNA1 Epstein-Barr virus protein by serine phosphorylation and arginine methylation. J Virol 80(11), 5261-72.

418. Sinclair, A. J., Palmero, I., Peters, G., and Farrell, P. J. (1994). EBNA-2 and EBNA-LP cooperate to cause G0 to G1 transition during immortalization of resting human B lymphocytes by Epstein-Barr virus. EMBO J 13(14), 3321-8.

419. Singh, G., Aras, S., Zea, A. H., Koochekpour, S., and Aiyar, A. (2009). Optimal transactivation by Epstein-Barr nuclear antigen 1 requires the UR1 and ATH1 domains. J Virol 83(9), 4227-35.

420. Singh, R., George, J., and Shukla, Y. (2010). Role of senescence and mitotic catastrophe in cancer therapy. Cell Div 5, 4.

421. Sivachandran, N., Cao, J. Y., and Frappier, L. (2010). Epstein-Barr Virus Nuclear Antigen 1 Hijacks the Host Kinase CK2 To Disrupt PML Nuclear Bodies. Journal of Virology 84(21), 11113-11123.

422. Sivachandran, N., Dawson, C. W., Young, L. S., Liu, F. F., Middeldorp, J., and Frappier, L. (2012a). Contributions of the Epstein-Barr Virus EBNA1 Protein to Gastric Carcinoma. Journal of Virology 86(1), 60-68.

423. Sivachandran, N., Sarkari, F., and Frappier, L. (2008). Epstein-Barr Nuclear Antigen 1 Contributes to Nasopharyngeal Carcinoma through Disruption of PML Nuclear Bodies. Plos Pathogens 4(10).

424. Sivachandran, N., Wang, X., and Frappier, L. (2012b). Functions of the Epstein-Barr virus EBNA1 protein in viral reactivation and lytic infection. J Virol 86(11), 6146-58.

425. Sixbey, J. W., and Yao, Q. Y. (1992). Immunoglobulin A-induced shift of Epstein-Barr virus tissue tropism. Science 255(5051), 1578-80.

426. Smith, D. W., and Sugden, B. (2013). Potential Cellular Functions of Epstein-Barr Nuclear Antigen 1 (EBNA1) of Epstein-Barr Virus. Viruses 5(1), 226-240.

427. Smith, P. R., de Jesus, O., Turner, D., Hollyoake, M., Karstegl, C. E., Griffin, B. E., Karran, L., Wang, Y., Hayward, S. D., and Farrell, P. J. (2000). Structure and coding content of CST (BART) family RNAs of Epstein-Barr virus. J Virol 74(7), 3082-92.

428. Snudden, D. K., Hearing, J., Smith, P. R., Grasser, F. A., and Griffin, B. E. (1994). EBNA-1, the major nuclear antigen of Epstein-Barr virus, resembles 'RGG' RNA binding proteins. EMBO J 13(20), 4840-7.

429. Somasiri, A., Nielsen, J. S., Makretsov, N., McCoy, M. L., Prentice, L., Gilks, C. B., Chia, S. K., Gelmon, K. A., Kershaw, D. B., Huntsman, D. G., McNagny, K. M., and Roskelley, C. D. (2004). Overexpression of the anti-adhesin podocalyxin is an independent predictor of breast cancer progression. Cancer Res 64(15), 5068-73.

430. Soyer, O. S., and O'Malley, M. A. (2013). Evolutionary systems biology: what it is and why it matters. Bioessays 35(8), 696-705.

431. Spanjaard, R. A., and van Duin, J. (1988). Translation of the sequence AGG-AGG yields 50% ribosomal frameshift. Proc Natl Acad Sci U S A 85(21), 7967-71.

432. Srivastava, D. B., Ethayathulla, A. S., Kumar, J., Somvanshi, R. K., Sharma, S., Dey, S., and Singh, T. P. (2007). Carbohydrate binding properties and carbohydrate induced conformational switch in sheep secretory glycoprotein (SPS-40): crystal structures of four complexes of SPS-40 with chitin-like oligosaccharides. J Struct Biol 158(3), 255-66.

433. Steinmetz, M. O., and Akhmanova, A. (2008). Capturing protein tails by CAP-Gly domains. Trends Biochem Sci 33(11), 535-45.

434. Stevenson, D., Charalambous, C., and Wilson, J. B. (2005). Epstein-Barr virus latent membrane protein 1 (CAO) up-regulates VEGF and TGF alpha concomitant with hyperlasia, with subsequent up-regulation of p16 and MMP9. Cancer Res 65(19), 8826-35.

435. Stevenson, L. F., Sparks, A., Allende-Vega, N., Xirodimas, D. P., Lane, D. P., and Saville, M. K. (2007). The deubiquitinating enzyme USP2a regulates the p53 pathway by targeting Mdm2. EMBO J 26(4), 976-86.

436. Storer, A. C., and Menard, R. (1994). Catalytic mechanism in papain family of cysteine peptidases. Methods Enzymol 244, 486-500.

437. Strahl, B. D., and Allis, C. D. (2000). The language of covalent histone modifications. Nature 403(6765), 41-5.

438. Su, W., Middleton, T., Sugden, B., and Echols, H. (1991). DNA looping between the origin of replication of Epstein-Barr virus and its enhancer site: stabilization of an origin complex with Epstein-Barr nuclear antigen 1. Proc Natl Acad Sci U S A 88(23), 10870-4.

439. Sumara, I., Quadroni, M., Frei, C., Olma, M. H., Sumara, G., Ricci, R., and Peter, M. (2007). A Cul3-based E3 ligase removes Aurora B from mitotic chromosomes, regulating mitotic progression and completion of cytokinesis in human cells. Dev Cell 12(6), 887-900.

440. Sun, W., Tan, X., Shi, Y., Xu, G., Mao, R., Gu, X., Fan, Y., Yu, Y., Burlingame, S., Zhang, H., Rednam, S. P., Lu, X., Zhang, T., Fu, S., Cao, G., Qin, J., and Yang, J. (2010). USP11 negatively regulates TNFalpha-induced NF-kappaB activation by targeting on IkappaBalpha. Cell Signal 22(3), 386-94.

441. Sun, Y. J., Chang, N. C., Hung, S. I., Chang, A. C., Chou, C. C., and Hsiao, C. D. (2001). The crystal structure of a novel mammalian lectin, Ym1, suggests a saccharide binding site. J Biol Chem 276(20), 17507-14.

442. Sung, N. S., Wilson, J., Davenport, M., Sista, N. D., and Pagano, J. S. (1994). Reciprocal regulation of the Epstein-Barr virus BamHI-F promoter by EBNA-1 and an E2F transcription factor. Mol Cell Biol 14(11), 7144-52.

443. Sutherland, T. E., Andersen, O. A., Betou, M., Eggleston, I. M., Maizels, R. M., van Aalten, D., and Allen, J. E. (2011). Analyzing airway inflammation with chemical biology: dissection of acidic mammalian chitinase function with a selective drug-like inhibitor. Chem Biol 18(5), 569-79.

444. Suzuki, K., Sugawara, N., Suzuki, M., Uchiyama, T., Katouno, F., Nikaidou, N., and Watanabe, T. (2002). Chitinases A, B, and C1 of Serratia marcescens 2170 produced by recombinant Escherichia coli: enzymatic properties and synergism on chitin degradation. Biosci Biotechnol Biochem 66(5), 1075-83.

445. Swanson, W. J., Yang, Z., Wolfner, M. F., and Aquadro, C. F. (2001). Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci U S A 98(5), 2509-14.

446. Szalkowski, A. M. (2012). Fast and robust multiple sequence alignment with phylogeny-aware gap placement. BMC Bioinformatics 13, 129.

447. Szekely, L., Selivanova, G., Magnusson, K. P., Klein, G., and Wiman, K. G. (1993). EBNA-5, an Epstein-Barr virus-encoded nuclear antigen, binds to the retinoblastoma and p53 proteins. Proc Natl Acad Sci U S A 90(12), 5455-9.

448. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39(Database issue), D561-8.

449. Takada, K., and Nanbo, A. (2001). The role of EBERs in oncogenesis. Semin Cancer Biol 11(6), 461-7.

450. Takimoto, T., Sato, H., Ogura, H., Tanaka, S., Masuda, K., Ishikawa, S., and Umeda, R. (1989). Differences in the ability of cells to fuse are mediated by strains of Epstein-Barr virus. Laryngoscope 99(10 Pt 1), 1075-80.

451. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28(10), 2731-9.

452. Tanner, J., Whang, Y., Sample, J., Sears, A., and Kieff, E. (1988). Soluble gp350/220 and deletion mutant glycoproteins block Epstein-Barr virus adsorption to lymphocytes. J Virol 62(12), 4452-64.

453. Tegel, H., Tourle, S., Ottosson, J., and Persson, A. (2010). Increased levels of recombinant human proteins with the Escherichia coli strain Rosetta(DE3). Protein Expr Purif 69(2), 159-67.

454. Tellam, J., Rist, M., Connolly, G., Webb, N., Fazou, C., Wang, F., and Khanna, R. (2007). Translation efficiency of EBNA1 encoded by lymphocryptoviruses influences endogenous presentation of CD8+ T cell epitopes. Eur J Immunol 37(2), 328-37.

455. Tellam, J., Sherritt, M., Thomson, S., Tellam, R., Moss, D. J., Burrows, S. R., Wiertz, E., and Khanna, R. (2001). Targeting of EBNA1 for rapid intracellular degradation overrides the inhibitory effects of the Gly-Ala repeat domain and restores CD8+ T cell recognition. J Biol Chem 276(36), 33353-60.

456. Tharanathan, R. N., and Kittur, F. S. (2003). Chitin--the undisputed biomolecule of great potential. Crit Rev Food Sci Nutr 43(1), 61-87.

457. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25(24), 4876-82.

458. Tierney, R. J., Kao, K. Y., Nagra, J. K., and Rickinson, A. B. (2011). Epstein-Barr virus BamHI W repeat number limits EBNA2/EBNA-LP coexpression in newly infected B cells and the efficiency of B-cell transformation: a rationale for the multiple W repeats in wild-type virus strains. J Virol 85(23), 12362-75.

459. Tjoelker, L. W., Gosting, L., Frey, S., Hunter, C. L., Trong, H. L., Steiner, B., Brammer, H., and Gray, P. W. (2000). Structural and functional definition of the human chitinase chitin-binding domain. J Biol Chem 275(1), 514-20.

460. Tobi, D. (2010). Designing coarse grained-and atom based-potentials for protein-protein docking. BMC Struct Biol 10, 40.

461. Tomii, K., Sawada, Y., and Honda, S. (2012). Convergent evolution in structural elements of proteins investigated using cross profile analysis. BMC Bioinformatics 13, 11.

462. Tomkinson, B., and Kieff, E. (1992). Use of second-site homologous recombination to demonstrate that Epstein-Barr virus nuclear protein 3B is not important for lymphocyte infection or growth transformation in vitro. J Virol 66(5), 2893-903.

463. Tomkinson, B., Robertson, E., and Kieff, E. (1993). Epstein-Barr virus nuclear proteins EBNA-3A and EBNA-3C are essential for B-lymphocyte growth transformation. J Virol 67(4), 2014-25.

464. Tompa, P. (2011). Unstructural biology coming of age. Curr Opin Struct Biol 21(3), 419-25.

465. Tovchigrechko, A., and Vakser, I. A. (2006). GRAMM-X public web server for protein-protein docking. Nucleic Acids Res 34(Web Server issue), W310-4.

466. Tsai, M. L., Liaw, S. H., and Chang, N. C. (2004). The crystal structure of Ym1 at 1.31 A resolution. J Struct Biol 148(3), 290-6.

467. Tsimbouri, P., Drotar, M. E., Coy, J. L., and Wilson, J. B. (2002). bcl-x(L) and RAG genes are induced and the response to IL-2 enhanced in E mu EBNA-1 transgenic mouse lymphocytes. Oncogene 21(33), 5182-5187.

468. Tsurumi, T., Fujita, M., and Kudoh, A. (2005). Latent and lytic Epstein-Barr virus replication strategies. Rev Med Virol 15(1), 3-15.

469. Tsutsumi, S., Ohga, S., Nomura, A., Takada, H., Sakai, S., Ohshima, K., Sumimoto, K., and Hara, T. (2002). CD4-CD8- T-cell polymyositis in a patient with chronic active Epstein-Barr virus infection. Am J Hematol 71(3), 211-5.

470. Tsvetkov, P., Reuven, N., and Shaul, Y. (2009). The nanny model for IDPs. Nat Chem Biol 5(11), 778-81.

471. Tugizov, S. M., Berline, J. W., and Palefsky, J. M. (2003). Epstein-Barr virus infection of polarized tongue and nasopharyngeal epithelial cells. Nat Med 9(3), 307-14.

472. Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41(3), 415-27.

473. Uversky, V. N., Gillespie, J. R., Millett, I. S., Khodyakova, A. V., Vasilenko, R. N., Vasiliev, A. M., Rodionov, I. L., Kozlovskaya, G. D., Dolgikh, D. A., Fink, A. L., Doniach, S., Permyakov, E. A., and Abramov, V. M. (2000). Zn(2+)-mediated structure formation and compaction of the "natively unfolded" human prothymosin alpha. Biochem Biophys Res Commun 267(2), 663-8.

474. Valentine, R., Dawson, C. W., Hu, C. F., Shah, K. M., Owen, T. J., Date, K. L., Maia, S. P., Shao, J., Arrand, J. R., Young, L. S., and O'Neil, J. D. (2010). Epstein-Barr virus-encoded EBNA1 inhibits the canonical NF-kappa B pathway in carcinoma cells by inhibiting IKK phosphorylation. Molecular Cancer 9.

475. van Aalten, D. M., Komander, D., Synstad, B., Gaseidnes, S., Peter, M. G., and Eijsink, V. G. (2001). Structural insights into the catalytic mechanism of a family 18 exo-chitinase. Proc Natl Acad Sci U S A 98(16), 8979-84.

476. van Beek, J., zur Hausen, A., Klein Kranenbarg, E., van de Velde, C. J., Middeldorp, J. M., van den Brule, A. J., Meijer, C. J., and Bloemena, E. (2004). EBV-positive gastric adenocarcinomas: a distinct clinicopathologic entity with a low frequency of lymph node involvement. J Clin Oncol 22(4), 664-70.

477. van Delft, M. F., Wei, A. H., Mason, K. D., Vandenberg, C. J., Chen, L., Czabotar, P. E., Willis, S. N., Scott, C. L., Day, C. L., Cory, S., Adams, J. M., Roberts, A. W., and Huang, D. C. (2006). The BH3 mimetic ABT-737 targets selective Bcl-2 proteins and efficiently induces apoptosis via Bak/Bax if Mcl-1 is neutralized. Cancer Cell 10(5), 389-99.

478. van Leuken, R. J., Luna-Vargas, M. P., Sixma, T. K., Wolthuis, R. M., and Medema, R. H. (2008). Usp39 is essential for mitotic spindle checkpoint integrity and controls mRNA-levels of aurora B. Cell Cycle 7(17), 2710-9.

479. Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., and Teichmann, S. A. (2004). Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol 14(2), 208-16.

480. Volkmer, R., Tapia, V., and Landgraf, C. (2012). Synthetic peptide arrays for investigating protein interaction domains. FEBS Lett 586(17), 2780-6.

481. Vucetic, S., Brown, C. J., Dunker, A. K., and Obradovic, Z. (2003). Flavors of protein disorder. Proteins 52(4), 573-84.

482. Wagner, G., Laine, L., and Almeder, M. (1993). Chitin in the epidermal cuticle of a vertebrate (Paralipophrys trigloides, Blenniidae, Teleostei). Experientia 49, 317-319.

483. Waltzer, L., Perricaudet, M., Sergeant, A., and Manet, E. (1996). Epstein-Barr virus EBNA3A and EBNA3C proteins both repress RBP-J kappa-EBNA2-activated transcription by inhibiting the binding of RBP-J kappa to DNA. J Virol 70(9), 5909-15.

484. Wang, C., Mavrogianis, P. A., and Fazleabas, A. T. (2009). Endometriosis is associated with progesterone resistance in the baboon (Papio anubis) oviduct: evidence based on the localization of oviductal glycoprotein 1 (OVGP1). Biol Reprod 80(2), 272-8.

485. Wang, D., Liebowitz, D., and Kieff, E. (1985). An EBV membrane protein expressed in immortalized lymphocytes transforms established rodent cells. Cell 43(3 Pt 2), 831-40.

486. Wang, F., Gregory, C., Sample, C., Rowe, M., Liebowitz, D., Murray, R., Rickinson, A., and Kieff, E. (1990a). Epstein-Barr virus latent membrane protein (LMP1) and nuclear proteins 2 and 3C are effectors of phenotypic changes in B lymphocytes: EBNA-2 and LMP1 cooperatively induce CD23. J Virol 64(5), 2309-18.

487. Wang, F., Tsang, S. F., Kurilla, M. G., Cohen, J. I., and Kieff, E. (1990b). Epstein-Barr virus nuclear antigen 2 transactivates latent membrane protein LMP1. J Virol 64(7), 3407-16.

488. Wang, J. C., P. Kollman, PA (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J Comput Chem 21, 1049-1074.

489. Wang, L., Du, F., and Wang, X. (2008). TNF-alpha induces two distinct caspase-8 activation pathways. Cell 133(4), 693-703.

490. Wang, S., and Frappier, L. (2009). Nucleosome assembly proteins bind to Epstein-Barr virus nuclear antigen 1 and affect its functions in DNA replication and transcriptional activation. J Virol 83(22), 11704-14.

491. Wang, X., Kenyon, W. J., Li, Q., Mullberg, J., and Hutt-Fletcher, L. M. (1998). Epstein-Barr virus uses different complexes of glycoproteins gH and gL to infect B lymphocytes and epithelial cells. J Virol 72(7), 5552-8.

492. Wang, X., Wang, R., Zhang, Y., and Zhang, H. (2013). Evolutionary survey of druggable protein targets with respect to their subcellular localizations. Genome Biol Evol 5(7), 1291-7.

493. Wang, Y., Finan, J. E., Middeldorp, J. M., and Hayward, S. D. (1997). P32/TAP, a cellular protein that interacts with EBNA-1 of Epstein-Barr virus. Virology 236(1), 18-29.

494. Watanabe, T., Kobori, K., Miyashita, K., Fujii, T., Sakai, H., Uchida, M., and Tanaka, H. (1993). Identification of glutamic acid 204 and aspartic acid 200 in chitinase A1 of Bacillus circulans WL-12 as essential residues for chitinase activity. J Biol Chem 268(25), 18567-72.

495. Weaver, P. G., Doguzhaeva, L. A., Lawver, D. R., Tacker, R. C., Ciampaglio, C. N., Crate, J. M., and Zheng, W. (2011). Characterization of organics consistent with beta-chitin preserved in the Late Eocene cuttlefish Mississaepia mississippiensis. PLoS One 6(11), e28195.

496. Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D., Schuler, G. D., Shumway, M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2008). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36(Database issue), D13-21.

497. Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18(5), 691-9.

498. White, R. E., Groves, I. J., Turro, E., Yee, J., Kremmer, E., and Allday, M. J. (2010). Extensive co-operation between the Epstein-Barr virus EBNA3 proteins in the manipulation of host gene expression and epigenetic chromatin modification. PLoS One 5(11), e13979.

499. Wilson, J. B., Bell, J. L., and Levine, A. J. (1996). Expression of Epstein-Barr virus nuclear antigen-1 induces B cell neoplasia in transgenic mice. EMBO J 15(12), 3117-26.

500. Wilson, J. B., and Levine, A. J. (1992). The oncogenic potential of Epstein-Barr virus nuclear antigen 1 in transgenic mice. Curr Top Microbiol Immunol 182, 375-84.

501. Wilson, J. B., Weinberg, W., Johnson, R., Yuspa, S., and Levine, A. J. (1990). Expression of the BNLF-1 oncogene of Epstein-Barr virus in the skin of transgenic mice induces hyperplasia and aberrant expression of keratin 6. Cell 61(7), 1315-27.

502. Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2(5), 333-41.

503. Wood, V. H. J., O'Neil, J. D., Wei, W., Stewart, S. E., Dawson, C. W., and Young, L. S. (2007). Epstein-Barr virus-encoded EBNA1 regulates cellular gene transcription and modulates the STAT1 and TGF beta signaling pathways. Oncogene 26(28), 4135-4147.

504. Wright, J., Falk, L., and Deinhardt, F. (1975). Appearance of Epstein-Barr virus nuclear antigen in human cordblood lymphocytes. IARC Sci Publ(11 Pt 1), 409-14.

505. Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C. L., Haase, J., Janes, J., Huss, J. W., 3rd, and Su, A. I. (2009). BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol 10(11), R130.

506. Wu, H., Kapoor, P., and Frappier, L. (2002). Separation of the DNA replication, segregation, and transcriptional activation functions of Epstein-Barr nuclear antigen 1. J Virol 76(5), 2480-90.

507. Wu, S., Skolnick, J., and Zhang, Y. (2007a). Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 5, 17.

508. Wu, S., and Zhang, Y. (2007b). LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 35(10), 3375-82.

509. Wu, Y., Maruo, S., Yajima, M., Kanda, T., and Takada, K. (2007). Epstein-Barr virus (EBV)-encoded RNA 2 (EBER2) but not EBER1 plays a critical role in EBV-induced B-cell growth transformation. J Virol 81(20), 11236-45.

510. Xu, M., Takanashi, M., Oikawa, K., Tanaka, M., Nishi, H., Isaka, K., Kudo, M., and Kuroda, M. (2009). USP15 plays an essential role for caspase-3 activation during Paclitaxel-induced apoptosis. Biochem Biophys Res Commun 388(2), 366-71.

511. Yajima, M., Kanda, T., and Takada, K. (2005). Critical role of Epstein-Barr Virus (EBV)-encoded RNA in efficient EBV-induced B-lymphocyte growth transformation. J Virol 79(7), 4298-307.

512. Yamamoto, K., Matsuo, T., and Osato, T. (1975). Appearance of Epstein-Barr virus-determined nuclear antigen in human epithelial cells following fusion with lymphoid cells. Intervirology 6(2), 115-21.

513. Yang, H. (2013). Conserved or lost: molecular evolution of the key gene GULO in vertebrate vitamin C biosynthesis. Biochem Genet 51(5-6), 413-25.

514. Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: principles and practice. Nat Rev Genet 13(5), 303-14.

515. Yates, J. L., Warren, N., and Sugden, B. (1985). Stable replication of plasmids derived from Epstein-Barr virus in various mammalian cells. Nature 313(6005), 812-5.

516. Yin, Q., and Flemington, E. K. (2006). siRNAs against the Epstein Barr virus latency replication factor, EBNA1, inhibit its function and growth of EBV-dependent tumor cells. Virology 346(2), 385-93.

517. Yin, Y., Manoury, B., and Fahraeus, R. (2003). Self-inhibition of synthesis and antigen presentation by Epstein-Barr virus-encoded EBNA1. Science 301(5638), 1371-4.

518. Yong, P., Gu, Z., Luo, J. P., Wang, J. R., and Tso, J. K. (2002). Antibodies against the C-terminal peptide of rabbit oviductin inhibit mouse early embryo development to pass 2-cell stage. Cell Res 12(1), 69-78.

519. Young, L. S., and Rickinson, A. B. (2004). Epstein-Barr virus: 40 years on. Nat Rev Cancer 4(10), 757-68.

520. Yu, J., Zhou, Y., Tanaka, I., and Yao, M. (2010). Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. Bioinformatics 26(1), 46-52.

521. Yuan, J., Luo, K., Zhang, L., Cheville, J. C., and Lou, Z. (2010). USP10 regulates p53 localization and stability by deubiquitinating p53. Cell 140(3), 384-96.

522. Yuasa-Kawada, J., Kinoshita-Kawada, M., Rao, Y., and Wu, J. Y. (2009). Deubiquitinating enzyme USP33/VDU1 is required for Slit signaling in inhibiting breast cancer cell migration. Proc Natl Acad Sci U S A 106(34), 14530-5.

523. Yuasa-Kawada, J., Kinoshita-Kawada, M., Wu, G., Rao, Y., and Wu, J. Y. (2009). Midline crossing and Slit responsiveness of commissural axons require USP33. Nat Neurosci 12(9), 1087-9.

524. Zacharias, M. (2010). Accounting for conformational changes during protein-protein docking. Curr Opin Struct Biol 20(2), 180-6.

525. Zhang, Y. (2009). Protein structure prediction: when is it useful? Curr Opin Struct Biol 19(2), 145-55.

526. Zhang, Y., Devries, M. E., and Skolnick, J. (2006). Structure modeling of all identified G protein-coupled receptors in the human genome. PLoS Comput Biol 2(2), e13.

527. Zhang, C., Vasmatzis, G., Cornette, J. L., and DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 267(3), 707-26.

528. Zhang, D., Zaugg, K., Mak, T. W., and Elledge, S. J. (2006). A role for the deubiquitinating enzyme USP28 in control of the DNA-damage response. Cell 126(3), 529-42.

529. Zhang, X. Y., Pfeiffer, H. K., Thorne, A. W., and McMahon, S. B. (2008). USP22, an hSAGA subunit and potential cancer stem cell marker, reverses the polycomb-catalyzed ubiquitylation of histone H2A. Cell Cycle 7(11), 1522-4.

530. Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9, 40.

531. Zhang, Y. (2009). I-TASSER: fully automated protein structure prediction in CASP8. Proteins 77 Suppl 9, 100-13.

532. Zhang, Y., Kihara, D., and Skolnick, J. (2002). Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 48(2), 192-201.

533. Zhang, Y., and Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 101(20), 7594-9.

534. Zhao, B., Marshall, D. R., and Sample, C. E. (1996). A conserved domain of the Epstein-Barr virus nuclear antigens 3A and 3C binds to a discrete domain of Jkappa. J Virol 70(7), 4228-36.

535. Zhao, J., Yeong, L. H., and Wong, W. S. (2007). Dexamethasone alters bronchoalveolar lavage fluid proteome in a mouse asthma model. Int Arch Allergy Immunol 142(3), 219-29.

536. Zhao, X., Fiske, B., Kawakami, A., Li, J., and Fisher, D. E. (2011). Regulation of MITF stability by the USP13 deubiquitinase. Nat Commun 2, 414.

537. Zhao, Z. M., Reynolds, A. B., and Gaucher, E. A. (2011). The evolutionary history of the catenin gene family during metazoan evolution. BMC Evol Biol 11, 198.

538. Zhou, J., Deng, Z., Norseen, J., and Lieberman, P. M. (2010). Regulation of Epstein-Barr virus origin of plasmid replication (OriP) by the S-phase checkpoint kinase Chk2. J Virol 84(10), 4979-87.

539. Zhou, J., Snyder, A. R., and Lieberman, P. M. (2009). Epstein-Barr virus episome stability is coupled to a delay in replication timing. J Virol 83(5), 2154-62.

540. Zhou, Y., and Mishra, B. (2005). Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. Proc Natl Acad Sci U S A 102(11), 4051-6.

541. Zhu, C. B., Chen, L. L., Tian, J. J., Su, L., Wang, C., Gai, Z. T., Du, W. J., and Ma, G. L. (2012a). Elevated serum YKL-40 level predicts poor prognosis in hepatocellular carcinoma after surgery. Ann Surg Oncol 19(3), 817-25.

542. Zhu, C. B., Wang, C., Chen, L. L., Ma, G. L., Zhang, S. C., Su, L., Tian, J. J., and Gai, Z. T. (2012b). Serum YKL-40 independently predicts outcome after transcatheter arterial chemoembolization of hepatocellular carcinoma. PLoS One 7(9), e44648.

543. Zhu, X., Menard, R., and Sulea, T. (2007). High incidence of ubiquitin-like domains in human ubiquitin-specific proteases. Proteins 69(1), 1-7.

544. Zhu, Z., Zheng, T., Homer, R. J., Kim, Y. K., Chen, N. Y., Cohn, L., Hamid, Q., and Elias, J. A. (2004). Acidic mammalian chitinase in asthmatic Th2 inflammation and IL-13 pathway activation. Science 304(5677), 1678-82.

545. Ziegler, A., Nervi, P., Durrenberger, M., and Seelig, J. (2005). The cationic cell-penetrating peptide CPP(TAT) derived from the HIV-1 protein TAT is rapidly transported into living fibroblasts: optical, biophysical, and metabolic evidence. Biochemistry 44(1), 138-48.

546. zur Hausen, A., Brink, A. A., Craanen, M. E., Middeldorp, J. M., Meijer, C. J., and van den Brule, A. J. (2000). Unique transcription pattern of Epstein-Barr virus (EBV) in EBV-carrying gastric adenocarcinomas: expression of the transforming BARF1 gene. Cancer Res 60(10), 2745-8.

Appendix I. Accession number of EBNA1 sequences

No.	Annotation	UniProt Acession Numbers
1	EBNA1 from EBV GD1	P03211
2	EBNA1 from EBV B958	Q3KSS4
3	EBNA1 from EBV AG876	Q1HVF7
4	EBNA1 from CyEBV TsBB6	Q9IPQ9
5	EBNA1 from CyEBV SiIIA	Q9IPQ8
6	EBNA1 from CeHV15	O91332
7	EBNA1 from CeHV12	Q80890
8	EBNA1 from CalHV3	Q993H1

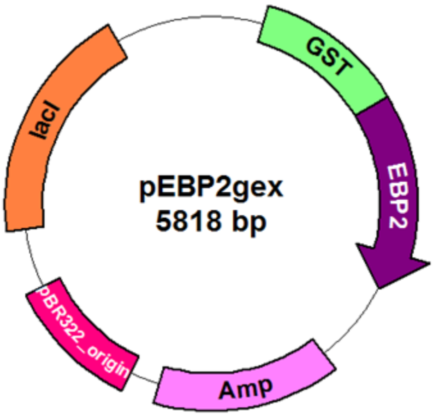
Appendix II. Peptide array sequences (analytical array)

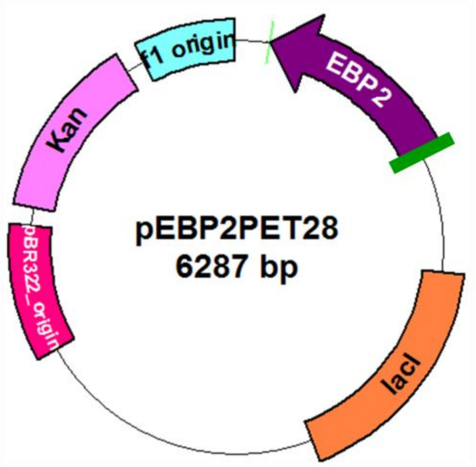
Spot No.	Sequence	Start	End
1	M-S-D-E-G-P-G-T-G-P-G-N-G-L-G-E-K-G-D-T-S-G-P-E-G	1	25
2	P-G-T-G-P-G-N-G-L-G-E-K-G-D-T-S-G-P-E-G-S-G-G-S-G	6	30
3	G-N-G-L-G-E-K-G-D-T-S-G-P-E-G-S-G-G-S-G-P-Q-R-R-G	11	35
4	E-K-G-D-T-S-G-P-E-G-S-G-G-S-G-P-Q-R-R-G-G-D-N-H-G	16	40
5	S-G-P-E-G-S-G-G-S-G-P-Q-R-R-G-G-D-N-H-G-R-G-R-G-R	21	45
6	S-G-G-S-G-P-Q-R-R-G-G-D-N-H-G-R-G-R-G-R-G-R-G-R-G	26	50
7	P-Q-R-R-G-G-D-N-H-G-R-G-R-G-R-G-R-G-R-G-G-G-R-P-G	31	55
8	G-D-N-H-G-R-G-R-G-R-G-R-G-R-G-G-G-R-P-G-A-P-G-G-S	36	60
9	R-G-R-G-R-G-R-G-R-G-G-G-R-P-G-A-P-G-G-S-G-S-G-P-R	41	65
10	G-R-G-R-G-G-G-R-P-G-A-P-G-G-S-G-S-G-P-R-H-R-D-G-V	46	70
11	G-G-R-P-G-A-P-G-G-S-G-S-G-P-R-H-R-D-G-V-R-R-P-Q-K	51	75
12	A-P-G-G-S-G-S-G-P-R-H-R-D-G-V-R-R-P-Q-K-R-P-S-C-I	56	80
13	G-S-G-P-R-H-R-D-G-V-R-R-P-Q-K-R-P-S-C-I-G-C-K-G-T	61	85
14	H-R-D-G-V-R-R-P-Q-K-R-P-S-C-I-G-C-K-G-T-H-G-G-T-G	66	90
15	R-R-P-Q-K-R-P-S-C-I-G-C-K-G-T-H-G-G-T-G-A-G-A-G-A	71	95
16	R-P-S-C-I-G-C-K-G-T-H-G-G-T-G-A-G-A-G-A-G-G-A-G-A	76	100
17	G-C-K-G-T-H-G-G-T-G-A-G-A-G-A-G-G-A-G-A-G-G-A-G-A	81	105
18	H-G-G-T-G-A-G-A-G-A-G-G-A-G-A-G-G-A-G-A-G-G-G-A-G	86	110
19	A-G-A-G-A-G-G-A-G-A-G-G-A-G-A-G-G-G-A-G-A-G-G-G-A	91	115
20	G-G-A-G-A-G-G-A-G-A-G-G-G-A-G-A-G-G-G-A-G-G-A-G-G	96	120
21	G-G-A-G-A-G-G-G-A-G-A-G-G-G-A-G-G-A-G-G-A-G-G-A-G	101	125
22	G-G-G-A-G-A-G-G-G-A-G-G-A-G-G-A-G-G-A-G-A-G-G-G-A	106	130
23	A-G-G-G-A-G-G-A-G-G-A-G-G-A-G-A-G-G-G-A-G-A-G-G-G	111	135
24	G-G-A-G-G-A-G-G-A-G-A-G-G-G-A-G-A-G-G-G-A-G-G-A-G	116	140
25	A-G-G-A-G-A-G-G-G-A-G-A-G-G-G-A-G-G-A-G-G-A-G-A-G	121	145
26	A-G-G-G-A-G-A-G-G-G-A-G-G-A-G-G-A-G-A-G-G-G-A-G-A	126	150
27	G-A-G-G-G-A-G-G-A-G-G-A-G-A-G-G-G-A-G-A-G-G-G-A-G	131	155
28	A-G-G-A-G-G-A-G-A-G-G-G-A-G-A-G-G-G-A-G-G-A-G-A-G	136	160
29	G-A-G-A-G-G-G-A-G-A-G-G-G-A-G-G-A-G-A-G-G-G-A-G-G	141	165
30	G-G-A-G-A-G-G-G-A-G-G-A-G-A-G-G-G-A-G-G-A-G-G-A-G	146	170
31	G-G-G-A-G-G-A-G-A-G-G-G-A-G-G-A-G-G-A-G-A-G-G-G-A	151	175
32	G-A-G-A-G-G-G-A-G-G-A-G-G-A-G-A-G-G-G-A-G-A-G-G-G	156	180
33	G-G-A-G-G-A-G-G-A-G-A-G-G-G-A-G-A-G-G-G-A-G-G-A-G	161	185
34	A-G-G-A-G-A-G-G-G-A-G-A-G-G-G-A-G-G-A-G-A-G-G-G-A	166	190
35	A-G-G-G-A-G-A-G-G-G-A-G-G-A-G-A-G-G-G-A-G-G-A-G-G	171	195
36	G-A-G-G-G-A-G-G-A-G-A-G-G-G-A-G-G-A-G-G-A-G-A-G-G	176	200
37	A-G-G-A-G-A-G-G-G-A-G-G-A-G-G-A-G-A-G-G-G-A-G-A-G	181	205
38	A-G-G-G-A-G-G-A-G-G-A-G-A-G-G-G-A-G-A-G-G-A-G-G-A	186	210
39	G-G-A-G-G-A-G-A-G-G-G-A-G-A-G-G-A-G-G-A-G-G-A-G-A	191	215
40	A-G-A-G-G-G-A-G-A-G-G-A-G-G-A-G-G-A-G-A-G-G-A-G-A	196	220
41	G-A-G-A-G-G-A-G-G-A-G-G-A-G-A-G-G-A-G-A-G-G-G-A-G	201	225

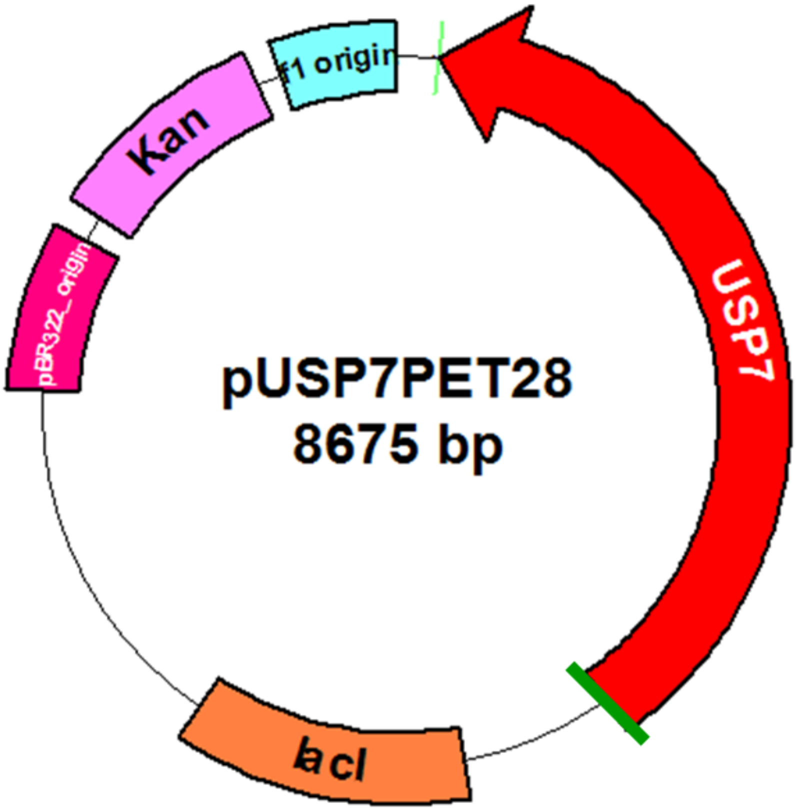
42	G-A-G-G-A-G-G-A-G-A-G-G-A-G-A-G-G-G-A-G-G-A-G-G-A	206	230
43	G-G-A-G-A-G-G-A-G-A-G-G-G-A-G-G-A-G-G-A-G-A-G-G-A	211	235
44	G-G-A-G-A-G-G-G-A-G-G-A-G-G-A-G-A-G-G-A-G-A-G-G-A	216	240
45	G-G-G-A-G-G-A-G-G-A-G-A-G-G-A-G-A-G-G-A-G-A-G-G-A	221	245
46	G-A-G-G-A-G-A-G-G-A-G-A-G-G-A-G-A-G-G-A-G-A-G-G-A	226	250
47	G-A-G-G-A-G-A-G-G-A-G-A-G-G-A-G-A-G-G-A-G-G-A-G-A	231	255
48	G-A-G-G-A-G-A-G-G-A-G-A-G-G-A-G-G-A-G-A-G-G-A-G-G	236	260
49	G-A-G-G-A-G-A-G-G-A-G-G-A-G-A-G-G-A-G-G-A-G-A-G-G	241	265
50	G-A-G-G-A-G-G-A-G-A-G-G-A-G-G-A-G-A-G-G-A-G-G-A-G	246	270
51	G-G-A-G-A-G-G-A-G-G-A-G-A-G-G-A-G-G-A-G-A-G-G-G-A	251	275
52	G-G-A-G-G-A-G-A-G-G-A-G-G-A-G-A-G-G-G-A-G-G-A-G-A	256	280
53	A-G-A-G-G-A-G-G-A-G-A-G-G-G-A-G-G-A-G-A-G-G-G-A-G	261	285
54	A-G-G-A-G-A-G-G-G-A-G-G-A-G-A-G-G-G-A-G-G-A-G-A-G	266	290
55	A-G-G-G-A-G-G-A-G-A-G-G-G-A-G-G-A-G-A-G-G-A-G-G-A	271	295
56	G-G-A-G-A-G-G-G-A-G-G-A-G-A-G-G-A-G-G-A-G-A-G-G-A	276	300
57	G-G-G-A-G-G-A-G-A-G-G-A-G-G-A-G-A-G-G-A-G-G-A-G-A	281	305
58	G-A-G-A-G-G-A-G-G-A-G-A-G-G-A-G-G-A-G-A-G-G-A-G-G	286	310
59	G-A-G-G-A-G-A-G-G-A-G-G-A-G-A-G-G-A-G-G-A-G-A-G-G	291	315
60	G-A-G-G-A-G-G-A-G-A-G-G-A-G-G-A-G-A-G-G-G-A-G-A-G	296	320
61	G-G-A-G-A-G-G-A-G-G-A-G-A-G-G-G-A-G-A-G-G-A-G-A-G	301	325
62	G-G-A-G-G-A-G-A-G-G-G-A-G-A-G-G-A-G-A-G-G-G-G-R-G	306	330
63	A-G-A-G-G-G-A-G-A-G-G-A-G-A-G-G-G-G-R-G-R-G-G-S-G	311	335
64	G-A-G-A-G-G-A-G-A-G-G-G-G-R-G-R-G-G-S-G-G-R-G-R-G	316	340
65	G-A-G-A-G-G-G-G-R-G-R-G-G-S-G-G-R-G-R-G-G-S-G-G-R	321	345
66	G-G-G-R-G-R-G-G-S-G-G-R-G-R-G-G-S-G-G-R-G-R-G-G-S	326	350
67	R-G-G-S-G-G-R-G-R-G-G-S-G-G-R-G-R-G-G-S-G-G-R-R-G	331	355
68	G-R-G-R-G-G-S-G-G-R-G-R-G-G-S-G-G-R-R-G-R-G-R-E-R	336	360
69	G-S-G-G-R-G-R-G-G-S-G-G-R-R-G-R-G-R-E-R-A-R-G-G-S	341	365
70	G-R-G-G-S-G-G-R-R-G-R-G-R-E-R-A-R-G-G-S-R-E-R-A-R	346	370
71	G-G-R-R-G-R-G-R-E-R-A-R-G-G-S-R-E-R-A-R-G-R-G-R-G	351	375
72	R-G-R-E-R-A-R-G-G-S-R-E-R-A-R-G-R-G-R-G-R-G-E-K-R	356	380
73	A-R-G-G-S-R-E-R-A-R-G-R-G-R-G-R-G-E-K-R-P-R-S-P-S	361	385
74	R-E-R-A-R-G-R-G-R-G-R-G-E-K-R-P-R-S-P-S-S-Q-S-S-S	366	390
75	G-R-G-R-G-R-G-E-K-R-P-R-S-P-S-S-Q-S-S-S-S-G-S-P-P	371	395
76	R-G-E-K-R-P-R-S-P-S-S-Q-S-S-S-S-G-S-P-P-R-R-P-P-P	376	400
77	P-R-S-P-S-S-Q-S-S-S-S-G-S-P-P-R-R-P-P-P-G-R-R-P-F	381	405
78	S-Q-S-S-S-S-G-S-P-P-R-R-P-P-P-G-R-R-P-F-F-H-P-V-G	386	410
79	S-G-S-P-P-R-R-P-P-P-G-R-R-P-F-F-H-P-V-G-E-A-D-Y-F	391	415
80	R-R-P-P-P-G-R-R-P-F-F-H-P-V-G-E-A-D-Y-F-E-Y-H-Q-E	396	420
81	G-R-R-P-F-F-H-P-V-G-E-A-D-Y-F-E-Y-H-Q-E-G-G-P-D-G	401	425
82	F-H-P-V-G-E-A-D-Y-F-E-Y-H-Q-E-G-G-P-D-G-E-P-D-V-P	406	430
83	E-A-D-Y-F-E-Y-H-Q-E-G-G-P-D-G-E-P-D-V-P-P-G-A-I-E	411	435
84	E-Y-H-Q-E-G-G-P-D-G-E-P-D-V-P-P-G-A-I-E-Q-G-P-A-D	416	440
85	G-G-P-D-G-E-P-D-V-P-P-G-A-I-E-Q-G-P-A-D-D-P-G-E-G	421	445
86	E-P-D-V-P-P-G-A-I-E-Q-G-P-A-D-D-P-G-E-G-P-S-T-G-P	426	450
87	P-G-A-I-E-Q-G-P-A-D-D-P-G-E-G-P-S-T-G-P-R-G-Q-G-D	431	455
88	Q-G-P-A-D-D-P-G-E-G-P-S-T-G-P-R-G-Q-G-D-G-G-R-R-K	436	460
89	D-P-G-E-G-P-S-T-G-P-R-G-Q-G-D-G-G-R-R-K-K-G-G-W-F	441	465
90	P-S-T-G-P-R-G-Q-G-D-G-G-R-R-K-K-G-G-W-F-G-K-H-R-G	446	470
91	R-G-Q-G-D-G-G-R-R-K-K-G-G-W-F-G-K-H-R-G-Q-G-G-S-N	451	475
92	G-G-R-R-K-K-G-G-W-F-G-K-H-R-G-Q-G-G-S-N-P-K-F-E-N	456	480
93	K-G-G-W-F-G-K-H-R-G-Q-G-G-S-N-P-K-F-E-N-I-A-E-G-L	461	485
94	G-K-H-R-G-Q-G-G-S-N-P-K-F-E-N-I-A-E-G-L-R-A-L-L-A	466	490
95	Q-G-G-S-N-P-K-F-E-N-I-A-E-G-L-R-A-L-L-A-R-S-H-V-E	471	495
96	P-K-F-E-N-I-A-E-G-L-R-A-L-L-A-R-S-H-V-E-R-T-T-D-E	476	500
97	I-A-E-G-L-R-A-L-L-A-R-S-H-V-E-R-T-T-D-E-G-T-W-V-A	481	505
98	R-A-L-L-A-R-S-H-V-E-R-T-T-D-E-G-T-W-V-A-G-V-F-V-Y	486	510
99	R-S-H-V-E-R-T-T-D-E-G-T-W-V-A-G-V-F-V-Y-G-G-S-K-T	491	515
100	R-T-T-D-E-G-T-W-V-A-G-V-F-V-Y-G-G-S-K-T-S-L-Y-N-L	496	520
101	G-T-W-V-A-G-V-F-V-Y-G-G-S-K-T-S-L-Y-N-L-R-R-G-T-A	501	525
102	G-V-F-V-Y-G-G-S-K-T-S-L-Y-N-L-R-R-G-T-A-L-A-I-P-Q	506	530

103	G-G-S-K-T-S-L-Y-N-L-R-R-G-T-A-L-A-I-P-Q-C-R-L-T-P	511	535
104	S-L-Y-N-L-R-R-G-T-A-L-A-I-P-Q-C-R-L-T-P-L-S-R-L-P	516	540
105	R-R-G-T-A-L-A-I-P-Q-C-R-L-T-P-L-S-R-L-P-F-G-M-A-P	521	545
106	L-A-I-P-Q-C-R-L-T-P-L-S-R-L-P-F-G-M-A-P-G-P-G-P-Q	526	550
107	C-R-L-T-P-L-S-R-L-P-F-G-M-A-P-G-P-G-P-Q-P-G-P-L-R	531	555
108	L-S-R-L-P-F-G-M-A-P-G-P-G-P-Q-P-G-P-L-R-E-S-I-V-C	536	560
109	F-G-M-A-P-G-P-G-P-Q-P-G-P-L-R-E-S-I-V-C-Y-F-M-V-F	541	565
110	G-P-G-P-Q-P-G-P-L-R-E-S-I-V-C-Y-F-M-V-F-L-Q-T-H-I	546	570
111	P-G-P-L-R-E-S-I-V-C-Y-F-M-V-F-L-Q-T-H-I-F-A-E-V-L	551	575
112	E-S-I-V-C-Y-F-M-V-F-L-Q-T-H-I-F-A-E-V-L-K-D-A-I-K	556	580
113	Y-F-M-V-F-L-Q-T-H-I-F-A-E-V-L-K-D-A-I-K-D-L-V-M-T	561	585
114	L-Q-T-H-I-F-A-E-V-L-K-D-A-I-K-D-L-V-M-T-K-P-A-P-T	566	590
115	F-A-E-V-L-K-D-A-I-K-D-L-V-M-T-K-P-A-P-T-C-N-I-R-V	571	595
116	K-D-A-I-K-D-L-V-M-T-K-P-A-P-T-C-N-I-R-V-T-V-C-S-F	576	600
117	D-L-V-M-T-K-P-A-P-T-C-N-I-R-V-T-V-C-S-F-D-D-G-V-D	581	605
118	K-P-A-P-T-C-N-I-R-V-T-V-C-S-F-D-D-G-V-D-L-P-P-W-F	586	610
119	C-N-I-R-V-T-V-C-S-F-D-D-G-V-D-L-P-P-W-F-P-P-M-V-E	591	615
120	T-V-C-S-F-D-D-G-V-D-L-P-P-W-F-P-P-M-V-E-G-A-A-A-E	596	620
121	D-D-G-V-D-L-P-P-W-F-P-P-M-V-E-G-A-A-A-E-G-D-D-G-D	601	625
122	L-P-P-W-F-P-P-M-V-E-G-A-A-A-E-G-D-D-G-D-D-G-D-E-G	606	630
123	P-P-M-V-E-G-A-A-A-E-G-D-D-G-D-D-G-D-E-G-G-D-G-D-E	611	635
124	G-A-A-A-E-G-D-D-G-D-D-G-D-E-G-G-D-G-D-E-G-E-E-G-Q	616	640
125	A-A-A-E-G-D-D-G-D-D-G-D-E-G-G-D-G-D-E-G-E-E-G-Q-E	621	645

Appendix III. Plasmid Maps







Appendix IV. Accession number of USPs

Homo sapiens				***Mus musculus***			
No.	**Accession No.**	**Given Annotation**	**Proposed Annotation**	**No.**	**Accession No.**	**Given Annotation**	**Proposed Annotation**
1	AAH50525	USP1	USP1	1	AAH20007	Usp1	Usp1
2	AAH02955	USP2	USP2	2	AAH17517	Usp2	Usp2
3	AAH18113	USP3	USP3	3	EDL26123	Usp3	Usp3
4	AAI25131	USP4	USP4	4	EDL21282	Usp4	Usp4
5	EAW88724	USP5	USP5	5	AAH66993	Usp5	Usp5
6	XP005256902	USP6	USP6	6	NP001003918	Usp7	Usp7
7	NP003461	USP7	USP7	7	AAH50947	Usp8	Usp8
8	AAI10591	USP8	USP8	8	P70398	Usp9x	Usp9x
9	XP005272732	USP9X	USP9X	9	NP_683745	Usp9y	Usp9y
10	EAW91608	USP9Y	USP9Y	10	EDL11619	Usp10	Usp10
11	AAH00263	USP10	USP10	11	EDL00740	Usp11	Usp11
12	AAI40850	USP11	USP11	12	AAH68136	Usp12	Usp12
13	AAH26072	USP12	USP12	13	AAH90999	Usp13	Usp13
14	AAH16146	USP13	USP13	14	AAH05571	Usp14	Usp14
15	AAH03556	USP14	USP14	15	AAH50042	Usp15	Usp15
16	EAW97104	USP15	USP15	16	AAH03278	Usp16	Usp16
17	EAX09927	USP16	USP16	17	NP001243902	Usp17	Usp17
18	NP958804	USP17	USP17	18	AAI38578	Usp18	Usp18
19	AAH14896	USP18	USP18	19	AAH60613	Usp19	Usp19
20	AAI46753	USP19	USP19	20	AAH79674	Usp20	Usp20
21	XP005251722	USP20	USP20	21	NP038947	Usp21	Usp21
22	AAH90946	USP21	USP21	22	AAH80737	Usp22	Usp22
23	NP_056091	USP22	USP22	23	NP899048	Usp24	Usp24
24	NP056121	USP24	USP24	24	AAH48171	Usp25	Usp25
25	AAH75792	USP25	USP25	25	AAK31949	Usp26	Usp26
26	AAK31972	USP26	USP26	26	NP062334	Usp27	Usp27
27	NP001138545	USP27	USP27	27	AAH88733	Usp28	Usp28
28	ACA06098	USP28	USP28	28	NP067298	Usp29	Usp29
29	NP065954	USP29	USP29	29	NP001028374	Usp30	Usp30
30	CAE51936	USP30	USP30	30	NP001028345	Usp31	Usp31
31	CAE51935	USP31	USP31	31	NP001025105	Usp32	Usp32
32	NP115971	USP32	USP32	32	EDL11924	Usp33	Usp33
33	EAX06371	USP33	USP33	33	NP001177330	Usp34	Usp34
34	NP055524	USP34	USP34	34	NP001170883	Usp35	Usp35
35	CAE51937	USP35	USP35	35	NP001028700	Usp36	Usp36
36	AAH71582	USP36	USP36	36	AAI39092	Usp37	Usp37
37	AAI33010	USP37	USP37	37	AAH54404	Usp38	Usp38
38	AAH68975	USP38	USP38	38	AAH26983	Usp39	Usp39
39	EAW99490	USP39	USP39	39	NP001185502	Usp40	Us40
40	XP005246145	USP40	USP40	40	AAI37853	Usp42	Usp42
41	Q3LFD5	USP41	USP41	41	NP776115	Usp43	Usp43
42	CAE53097	USP42	USP42	42	NP001193780	Usp44	Usp44
43	AAI44042	USP43	USP43	43	AAH27768	Usp45	Usp45
44	AAH30704	USP44	USP44	44	AAH39916	Usp46	Usp46
45	CAE47746	USP45	USP45	45	NP796223	Usp47	Usp47
46	AAH37574	USP46	USP46	46	NP570949	Usp48	Usp48
47	XP005253054	USP47	USP47	47	AAH60712	Usp49	Usp49
48	XP005246063	USP48	USP48	48	AAH61020	Usp50	Usp50
49	CAE51939	USP49	USP49	49	NP001131019	Usp51	Usp51
50	CAE47745	USP50	USP50	50	AAH75686	Usp52	Usp52
51	CAE47750	USP51	USP51	51	AAI32340	Usp53	Usp53
52	NP001120932	USP52	USP52	52	NP084456	Usp54	Usp54
53	XP005263130	USP53	USP53	53	NP775545	Cyld	Cyld
54	NP689799	USP54	USP54				
55	XP005255868	CYLD	CYLD				

Bos taurus				*Canis lupus familaris*			
No.	**Accession No.**	**Given Annotation**	**Proposed Annotation**	**No.**	**Accession No.**	**Given Annotation**	**Proposed Annotation**
1	XP005899914	Usp1	Usp1	1	XP852320	Usp1	Usp1
2	AAI12867	Usp2	Usp2	2	XP852320	Usp2	Usp2
3	XP005211741	Usp3	Usp3	3	XP544715	Usp3	Usp3
4	DAA16901	Usp4	Usp4	4	XP003432923	Usp4	Usp4
5	NP001178985	Usp5	Usp5	5	XP543845	Usp5	Usp5
6	XP005904435	Usp7	Usp7	6	XP005621615	Usp7	Usp7
7	NP001069594	Usp8	Usp8	7	XP535474	Usp8	Usp8
8	XP002700252	Usp9x	Usp9x	8	XP005642140	Usp9x	Usp9x
9	NP001138981	Usp9y	Usp9y	9	AGS47768	Usp9y	Usp9y
10	DAA20297	Usp10	Usp10	10	XP005620940	Usp10	Usp10
11	XP005228137	Usp11	Usp11	11	NP001183969	Usp11	Usp11
12	ABQ13036	Usp12	Usp12	12	XP543159	Usp12	Usp12
13	DAA33286	Usp13	Usp13	13	XP003434175	Usp13	Usp13
14	AAI22667	Usp14	Usp14	14	XP537306	Usp14	Usp14
15	AAI05522	Usp15	Usp15	15	XP849935	Usp15	Usp15
16	AAI23862	Usp16	Usp16	16	XP848330	Usp16	Usp16
17	XP005196582	Usp17	Usp17	17	XP854031	Usp17	Usp17
18	DAA29462	Usp18	Usp18	18	XP005637457	Usp18	Usp18
19	XP005196535	Usp19	Usp19	19	XP005632614	Usp19	Usp19
20	DAA24140	Usp20	Usp20	20	XP005625298	Usp20	Usp20
21	AAI05489	Usp21	Usp21	21	XP536136	Usp21	Usp21
22	DAA18687	Usp22	Usp22	22	XP005620237	Usp22	Usp22
23	DAA31228	Usp24	Usp24	23	XP005620394	Usp24	Usp24
24	DAA33643	Usp25	Usp25	24	XP535562	Usp25	Usp25
25	DAA13384	Usp26	Usp26	25	XP005641887	Usp26	Usp26
26	NP001138547	Usp27	Usp27	26	ENSCAFG00000015958	Usp27	Usp27
27	DAA22406	Usp28	Usp28	27	XP005619814	Usp28	Usp28
28	XP005911409	Usp29	Usp29	28	XP005636387	Usp30	Usp30
29	DAA20724	Usp30	Usp30	29	XP005622126	Usp31	Usp31
30	XP002703139	Usp31	Usp31	30	XP537710	Usp32	Usp32
31	DAA19082	Usp32	Usp32	31	XP005622106	Usp33	Usp33
32	DAA31326	Usp33	Usp33	32	XP005626192	Usp34	Usp34
33	DAA24665	Usp34	Usp34	33	XP542286	Usp35	Usp35
34	XP002699105	Usp35	Usp35	34	XP005624096	Usp36	Usp36
35	XP580726	Usp36	Usp36	35	XP545643	Usp37	Usp37
36	DAA32431	Usp37	Usp37	36	XP533279	Usp38	Usp38
37	DAA20849	Usp38	Usp38	37	XP532977	Usp39	Usp39
38	DAA24597	Usp39	Usp39	38	XP005635930	Usp40	Us40
39	DAA30943	Usp40	Us40	39	XP005621173	Usp42	Usp42
40	XP005225215	Usp42	Usp42	40	XP005620104	Usp43	Usp43
41	DAA18798	Usp43	Usp43	41	XP532654	Usp44	Usp44
42	XP005206168	Usp44	Usp44	42	XP539054	Usp45	Usp45
43	XP005210908	Usp45	Usp45	43	XP005628243	Usp46	Usp46
44	NP001179373	Usp46	Usp46	44	XP005633729	Usp47	Usp47
45	NP001230219	Usp47	Usp47	45	XP535372	Usp48	Usp48
46	XP003581941	Usp48	Usp48	46	XP532134	Usp49	Usp49
47	XP005192988	Usp49	Usp49	47	XP850913	Usp50	Usp50
48	NP001073699	Usp50	Usp50	48	XP531635	Usp52	Usp52
49	XP005206706	Usp52	Usp52	49	XP005639380	Usp53	Usp53
50	XP003582353	Usp53	Usp53	50	XP005619079	Usp54	Usp54
51	XP003588038	Usp54	Usp54	51	XP005617624	Cyld	Cyld
52	XP005218740	Cyld	Cyld				

Monodelphis domesticus				*Ornithorhynchus anatinus*			
No.	**Accession No.**	**Given Annotation**	**Proposed Annotation**	**No.**	**Accession No.**	**Given Annotation**	**Proposed Annotation**
1	XP001380891	Usp1	Usp1	8	ENSOANG00000003109	Usp11	Usp11
2	XP001380891	Usp2	Usp2	9	XP001519444	Usp12	Usp12
3	XP001366252	Usp3	Usp3	10	XP001519267	Usp14	Usp14
4	XP001367947	Usp4	Usp4	11	XP001519648	Usp15	Usp15
5	XP001370137	Usp5	Usp5	12	XP003430360	Usp19	Usp19
6	XP003341676	Usp7	Usp7	13	XP001507999	Usp20	Usp20
7	XP001370028	Usp8	Usp8	14	XP001511206	Usp22	Usp22
8	XP001366553	Usp9	Usp9	15	XP001514451	Usp25	Usp25
9	XP001371319	Usp11	Usp11	16	XP001518118	Usp28	Usp28
10	XP001376012	Usp12	Usp12	17	XP001508127	Usp30	Usp30
11	XP001368216	Usp13	Usp13	18	XP001509900	Usp31	Usp31
12	XP001367917	Usp14	Usp14	19	XP001510553	Usp32	Usp32
13	XP001363243	Usp15	Usp15	20	XP001507122	Usp33	Usp33
14	XP001373230	Usp16	Usp16	21	XP001512478	Usp34	Usp34
15	XP001374147	Usp18	Usp18	22	XP001520693	Usp36	Usp36
16	XP001367829	Usp19	Usp19	23	XP001515367	Usp37	Usp37
17	XP001364410	Usp20	Usp20	24	XP001513216	Usp38	Usp38
18	XP001371948	Usp21	Usp21	25	XP001518784	Usp39	Usp39
19	XP001370855	Usp22	Usp22	26	XP001510790	Usp40	Usp40
20	XP003340143	Usp24	Usp24	27	ENSOANG00000003278	Usp42	Usp42
21	XP001381253	Usp28	Usp28	28	XP001510004	---	Usp44
22	XP003342245	Usp30	Usp30	29	XP001506389	Usp45	Usp45
23	XP001377884	Usp31	Usp31	30	XP001515746	Usp46	Usp46
24	ENSMODG00000014128	Usp32	Usp32	31	XP001511271	Usp47	Usp47
25	XP001382154	Usp34	Usp34	32	XP001510634	Usp48	Usp48
26	XP003340998	Usp35	Usp35	33	XP001518088	Usp49	Usp49
27	XP001371102	Usp36	Usp36	34	XP001506997	Usp50	Usp50
28	XP001365238	Usp37	Usp37	35	XP001513697	Usp52	Usp52
29	XP001367345	Usp38	Usp38	36	XP001512481	Usp53	Usp53
30	XP001363937	Usp39	Usp39	37	XP001520353	Usp54	Usp54
31	XP001376325	Usp40	Us40	38	XP003430149	Cyld	Cyld
32	XP001377522	Usp42	Usp42	*Anolis carolinensis*			
33	XP003340388	Usp43	Usp43	**No.**	**Accession No.**	**Given Annotation**	**Proposed Annotation**
34	XP001367818	Usp44	Usp44	1	XP003220186	usp1	usp1
35	XP001367818	Usp45	Usp45	2	XP003229751	usp2	usp2
36	XP001371592	Usp46	Usp46	3	XP003228123	usp3	usp3
37	XP001379670	Usp47	Usp47	4	XP003217654	usp4	usp4
38	XP001377884	Usp48	Usp48	5	XP003227153	usp5	usp5
39	XP001379917	Usp49	Usp49	6	XP003224810	usp7	usp7
40	XP001380549	Usp50	Usp50	7	XP003220426	usp8	usp8
41	XP001371690	Usp53	Usp53	8	XP003218990	FAF-X	usp9
42	XP001364881	Usp54	Usp54	9	XP003228345	usp10	usp10
43	XP001363603	Cyld	Cyld	10	ENSACAG00000004813	usp11	usp11
Ornithorhynchus anatinus							
No.	**Accession No.**	**Given Annotation**	**Proposed Annotation**	11	XP003225554	usp12	usp12
1	XP001514295	Usp2	Usp2	12	XP003218164	usp13	usp13
2	XP003429803	Usp3	Usp3	13	XP003219701	usp14	usp14
3	XP001505377	Usp4	Usp4	14	XP003221229	ups15	ups15
4	XP001506396	Usp7	Usp7	15	XP003219113	usp16	usp16
5	XP001507996	Usp8	Usp8	16	XP003220892	usp18	usp18
6	XP003430712	Faf-X	Usp9	17	XP003217915	usp19	usp19
7	XP001510486	Usp10	Usp10	18	XP003230291	usp20	usp20

Anolis carolinensis				*Xenopus tropicalis*			
No.	**Accession No.**	**Given Annotation**	**Proposed Annotation**	**No.**	**Accession No.**	**Given Annotation**	**Proposed Annotation**
19	ENSACAG00000004099	usp21	usp21	26	NP001016228	usp33	usp33
20	XP003226563	usp22	usp22	27	XP002939554	usp34	usp34
21	XP003220199	usp24	usp24	28	XP002936594	usp35	usp35
22	XP003219106	usp25	usp25	29	XP004916402	usp36	usp36
23	XP003226222	usp28	usp28	30	XP004917686	usp37	usp37
24	ENSACAG00000025026	usp30	usp30	31	XP004911187	usp38	usp38
25	XP003229499	usp31	usp31	32	XP004912614	usp39	usp39
26	XP003226642	usp32	usp32	33	XP002932051	usp40	usp40
27	XP003223123	usp33	usp33	34	XP004918037	usp42	usp42
28	XP003227816	usp34	usp34	35	XP004918792	usp43	usp43
29	XP003226021	usp35	usp35	36	NP001072389	usp44	usp44
30	XP003217171	usp36	usp36	37	NP001011153	usp45	usp45
31	G1KAT4	usp37	usp37	38	NP001106637	usp46	usp46
32	XP003221712	usp38	usp38	39	NP001090710	usp47	usp47
33	XP003228282	usp39	usp39	40	NP001120167	usp48	usp48
34	XP003215223	usp40	usp40	41	XP002933022	usp49	usp49
35	XP003227691	usp42	usp42	42	XP004911946	usp52	usp52
36	XP003217208	usp43	usp43	43	XP002934301	usp53	usp53
37	XP003221118	usp44	usp44	44	XP002935811	usp54	usp54
38	XP003215560	usp45	usp45	45	NP001116960	cyld	cyld
39	XP003225359	usp46	usp46	*Danio rerio*			
40	XP003224231	usp47	usp47		NP955873	usp1	usp1
41	XP003230296	usp48	usp48		XP001337596	usp2	usp2
42	XP003220426	usp49	usp49		NP001186800	usp3	usp3
43	XP003228929	usp50	usp50		XP002662556	usp4	usp4
44	XP003216988	usp52	usp52		XP005173535	usp5	usp5
45	XP003221824	usp53	usp53		XP005164014	usp7	usp7
46	XP003223182	usp54	usp54		XP005170811	usp8	usp8
Xenopus tropicalis					NP001070917	faf-x	usp9
1	NP001072581	usp1	usp1		XP685621	usp10	usp10
2	NP001135522	usp2	usp2		XP002663119	usp11	usp11
3	NP001006783	usp3	usp3		NP001077025	usp12	usp12
4	XP002936515	usp4	usp4		XP005165987	usp13	usp13
5	NP001116956	usp5	usp5		NP956267	usp14	usp14
6	XP002939495	usp7	usp7		XP002667650	usp15	usp15
7	XP004912670	usp8	usp8		NP001139569	usp16	usp16
8	ENSXETG00000015489	usp9	usp9		XP002661398	usp18	usp18
9	NP001006761	usp10	usp10		XP005162175	usp19	usp19
10	XP002941584	usp12	usp12		NP957281	usp20	usp20
11	XP002931622	usp13	usp13		XP692003	usp21	usp21
12	NP001005641	usp14	usp14		NP001038713	usp22	usp22
13	NP001121498	usp15	usp15		XP005170208	usp24	usp24
14	NP001072158	usp16	usp16		NP001001886	usp25	usp25
15	XP004912551	usp18	usp18		XP001920096	usp28	usp28
16	NP001072879	usp19	usp19		XP005165213	usp30	usp30
17	NP001090641	usp20	usp20		XP005164285	usp31	usp31
18	XP002942482	usp21	usp21		XP005157663	usp32	usp32
19	NP001192175	usp22	usp22		NP998392	usp33	usp33
20	XP002931653	usp24	usp24		XP002660609	usp34	usp34
21	NP001039152	usp25	usp25		XP688241	usp36	usp36
22	XP002937867	usp28	usp28		XP005169174	usp37	usp37
23	ENSXETG00000022301	usp30	usp30		XP003197722	usp38	usp38
24	XP002932037	usp31	usp31		NP001073539	usp39	usp39
25	XP002933869	usp32	usp32		XP001921353	usp40	usp40

Danio rerio				*Strongylocentrotus purpuratus*			
No.	**Accession No.**	**Given Annotation**	**Proposed Annotation**	**No.**	**Accession No.**	**Given Annotation**	**Proposed Annotation**
	XP005169521	usp42	usp42		XP790411	usp16	usp16/45
	NP001082871	usp43	usp43		XP003723365	usp19	usp19
	NP956551	usp44	usp44		XP792380	usp20	usp20/33
	XP005158432	usp45	usp45		XP003728926	usp20	usp20
	NP001231910	usp46	usp46		XP786312	usp22	usp22
	NP001093619	usp47	usp47		XP796637	usp24	usp24
	XP005172974	usp48	usp48		XP003723638	usp25	usp25/28
	NP001038361	usp49	usp49		XP785002	usp30	usp30
	XP001920000	usp52	usp52		XP798688	usp31	usp31/43
	XP005171014	usp53	usp53		XP003726385	usp32	usp32
	XP005156922	usp54	usp54		XP790530	usp34	usp34
	XP684817	cyld	cyld		XP003731526	----	usp36/42
	XP001334225	usp64e	usp64e		XP001179436	----	usp36/42
	XP003198930	usp17	usp37		XP795476	usp36	usp36/42
Ciona intestinalis					XP003725936	usp37	usp37
	XP002129026	usp2	usp2/21		XP001178896	usp38	usp35/38
	XP002128421	---	usp4/11/15		XP001185686	usp39	usp39
	XP002122471	usp5	usp5/13		XP782883	usp40	usp40
	XP004225844	usp7	usp7		XP792596	usp44	usp44/49
	XP002126339	---	usp8		XP003728913	usp44	usp44/49
	XP002123585	faf-x	usp9		XP001200360	usp47	usp47
	XP002120858	usp10	usp10		XP782863	usp47	usp47
	XP002120616	usp12	usp12/46		XP791204	usp48	usp48
	XP002126675	usp14	usp14		XP790587	usp52	usp52
	XP004226714	usp16	usp16/45		XP781136	---	usp53/54
	NP001071926	znf	usp19		XP782657	-----	cyld
	XP002124776	usp20	usp20/33		XP003726771	cyld	cyld
	XP004225789	usp22	usp22		XP788815	---	usp4/11/15
	XP002123370	usp25	usp25/28	*Drosophila melanogaster*			
	XP002131158	usp30	usp30	1	NP733282	cg15817	usp1
	XP002120550	usp31	usp31/43	2	NP608462	cg14619	usp2/21
	XP002125579	usp32	usp32	3	NP647773	cg12082	usp5/13
	XP002124833	usp33	usp20/33	4	NP572779	usp7	usp7
	XP002127053	usp36	usp36/42	5	NP650948	ubpy	usp8
	XP002127688	usp37	usp37	6	NP524612	fat facets	usp9
	XP002124170	usp38	usp35/38	7	NP728554	cg32479	usp10
	XP002123896	usp39	usp39	8	NP651099	cg7023	usp12/46
	XP002121238	usp40	usp40	9	NP609377	cg5384	usp14
	NP001041464	znf	usp44/49	10	NP572220	cg4165	usp16/45
	XP002128614	usp45	usp16/45	11	NP610943	cg8494	usp20
	XP002122214	usp47	usp47	12	NP524140	non stop	usp22
	XP002121467	usp48	usp48	13	NP572274	cg3016	usp30
	XP002122964	usp52	usp52	14	NP611959	cg30421	usp31
	XP002120975	usp54	usp53/54	15	NP649153	cg8334	usp32
	XP002131459	cyld	cyld	16	NP651275	cg5794	usp34
Strongylocentrotus purpuratus				17	NP610784	cg8830	usp35/38
	XP782306	usp1	usp1	18	NP729092	scrawny	usp36
	XP003728306	usp2	usp2	19	Np573334	cg7288	usp39
	XP781718	usp3	usp3	20	NP996001	64e	usp47
	XP003725951	usp4/11/15	usp4/11/15	21	NP610427	cg8232	usp52
	XP796964	usp5/13	usp5/13	22	Np570018	cg2662	usp53/54
	XP780569	usp7	usp7	23	NP723554	cyld	cyld
	XP784858	usp8	usp8	*Caenorhabditis elegans*			
	XP003723719	usp9	usp9	1	NP493434	usp3	usp3
	XP794239	usp10	usp10	2	NP501035	h34c03	usp4
	XP783431	usp12	usp12/46	3	NP491765	usp5	usp5/13
	XP786966	usp14	usp14	4	NP505825	math-33	usp7

Caenorahbidtis elegans			
No.	**Accession No.**	**Given Annotation**	**Proposed Annotation**
5	NP507513	e01b7	usp8
6	NP497006	usp14	usp14
7	NP001254304	f07a11	usp19
8	NP504537	c04e6	usp22
9	NP495932	t24b8	usp24
10	NP495696	k02c4	usp25/28
11	NP495213	h12i13	usp25/28
12	NP497422	y67d2	usp30
13	NP001022992	cyk3	usp32
14	NP510570	usp33	usp33
15	NP494298	usp39	usp39
16	NP499162	usp46	usp46
17	NP495686	t05h10	usp47
18	NP492524	usp48	usp48
19	NP498519	panl2	usp52
20	NP001255047	cyld	cyld
Hydra magnipapillata			
1	XP004208370	usp2	usp2
2	XP002167244	usp3	usp3
3	XP002168139	usp5	usp5
4	XP002166763	usp7	usp7
5	XP002164031	----	usp8
6	XP004209945	usp9	usp8
7	XP002160338	faf-x	usp9
8	XP002158653	usp10	usp10
9	XP002167422	usp11	usp
10	XP002165294	usp12	usp12/46
11	XP002162819	usp14	usp14
12	XP002169084	usp15	usp15
13	XP002156339	usp19	usp19
14	XP004212543	usp20	usp20
15	XP002159615	---	usp22
16	XP002156309	usp24	usp24
17	XP002154564	usp32	usp32
18	XP004211619	usp32	usp32
19	XP002169046	usp32	usp32
20	XP004206986	usp33	usp33
21	XP002159948	usp34	usp34
22	XP002169312	usp35	usp35
23	XP002157785	usp36	usp36
24	XP002162018	---	usp37
25	XP002155067	usp39	usp39
26	XP002163722	---	usp45
27	XP004207299	usp64e	usp47
28	XP002154983	usp48	usp48
29	XP002160677	usp52	usp52
Disctyostelium discoideum			
1	XP643147	---	usp7
2	XP643687	---	usp34
3	XP643982	c19	usp36

Dictyostelium discoideum			
No.	**Accession No.**	**Given Annotation**	**Proposed Annotation**
4	XP636128	c19	usp40
5	XP629788	c19	usp9
6	XP635785	atpase	usp12/46
7	XP640063	CHO binding	no human homologue
8	XP643907	usp48	usp48
9	XP644261	c19	usp16/45
10	XP628978	c19	usp4/11/15
11	XP001733062	c19	usp4/11/15
12	XP645628	c19	usp35/38
13	XP643807	uhb	usp39
14	XP647829	sap	usp10
15	XP645688	udcp	usp14
16	XP646784	c19	nh
Chlamydomonas reinhardtii			
1	XP001697158	----	usp14
2	XP001700098	uch	usp7
3	XP001689583	----	usp48
4	XP001696423	----	usp12/46
5	XP001692784	----	usp36/42
6	XP001702430	----	8/4/11/15
7	XP001696552	----	usp2
8	XP001700667	----	usp10
9	XP001696671	----	nh
10	XP001701682	----	usp5/13
11	XP001697206	----	usp39
12	XP001691395	usp52	usp52
13	XP001691026	----	nh

Miscellaneous Sequences

Bos mutus			
1	XP00591140	Usp29	9Usp29
Gasterosteus aculeatus			
1	ENSGACG00000015008	usp50	usp50
2	ENSGACG00000018996	usp15	usp15
3	ENSGACG00000002025	usp9	usp9
4	ENSGACG00000015728	usp47	usp47
Latimeria chalumnae			
1	ENSLACG00000001725	usp35	usp35
Sarcophilus harissi			
1	ENSSHAG00000001467	Usp33	Usp33
2	ENSSHAG00000006863	Usp10	Usp10
3	ENSSHAG00000006133	Usp25	Usp25
Sus scrofa			
1	XP003135158	Usp51	Usp51

Appendix V. Molecular partners of USPs

Symbol	Full name
AKT1	v-akt murine thymoma viral oncogene homolog 1
ARRB2	arrestin, beta 2
ATAD5	ATPase family, AAA domain containing 5
ATG3	autophagy related 3
ATXN1	ataxin 1
ATX7NL3	ataxin 7-like 3
BCL3	B-cell CLL/lymphoma 3
BMI1	BMI1 polycomb ring finger oncogene
BIRC2	baculoviral IAP repeat containing 2
BIRC5	baculoviral IAP repeat containing 5
BIRC6	baculoviral IAP repeat containing 6
BRCA1	breast cancer 1, early onset
BRCA2	breast cancer 2, early onset
BTBD9	BTB (POZ) domain containing 9
C10orf2	chromosome 10 open reading frame 2
C14orf2	chromosome 14 open reading frame 2
C18orf2	chromosome 18 open reading frame 2
CASP1	caspase 1
CBX8	chromobox homolog 8
CDK4	cyclin-dependent kinase 4
CDKL2	cyclin-dependent kinase-like 2
CFTR	cystic fibrosis transmembrane conductance regulator
CHMP1A	charged multivesicular body protein 1A
CHMP2A	charged multivesicular body protein 2A
CHMP4A	charged multivesicular body protein 4A
CHMP4C	charged multivesicular body protein 4C
CHMP6	charged multivesicular body protein 6
CLSPN	claspin
CTNNB1	catenin (cadherin-associated protein), beta 1
DAP6	death-domain associated protein
DAXX	death-domain associated protein
DDX58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58
DIO2	deiodinase, iodothyronine, type II
DMWD	dystrophia myotonica, WD repeat containing
DNAH5	dynein, axonemal, heavy chain 5
DNAH12	dynein, axonemal, heavy chain 12
DNAH14	dynein, axonemal, heavy chain 14
DNAH19	dynein, axonemal, heavy chain 19
DNAJB6	DnaJ (Hsp40) homolog, subfamily B, member 6
DNMT1	DNA (cytosine-5-)-methyltransferase 1
DVL1	dishevelled, dsh homolog 1
DYNC1H1	dynein, cytoplasmic 1, heavy chain 1
EGFR	epidermal growth factor receptor
EIF3CL	eukaryotic translation initiation factor 3, subunit C-like
ENY2	enhancer of yellow 2 homolog
EPS15	epidermal growth factor receptor pathway substrate 15
ERG	v-ets erythroblastosis virus E26 oncogene homolog
FANCD2	Fanconi anemia, complementation group D2
FANCI	Fanconi anemia, complementation group I
FASN	fatty acid synthase
FAM48A	suppressor of Ty 20 homolog
FBXL3	F-box and leucine-rich repeat protein 3
FBXL7	F-box and leucine-rich repeat protein 7
FBXL15	F-box and leucine-rich repeat protein 15
FBXO2	F-box protein 2
FBXW7	F-box and WD repeat domain containing 7, E3 ubiquitin protein ligase
FOXO4	forkhead box O4

FUCA1	fucosidase, alpha-L- 1, tissue
G3BP1	GTPase activating protein binding protein 1
G6PI	glucose-6-phosphate isomerase
G6PD	glucose-6-phosphate dehydrogenase
GMPS	guanine monphosphate synthetase
GRP	gastrin-releasing peptide
GRB2	growth factor receptor-bound protein 2
H2A	Histone 2A
H2B	Histone 2B
HAT	histone acetyltransferase
HDAC6	histone deacetylase 6
HUWE1	HECT, UBA and WWE domain containing 1 E3 ubiquitin ligase
IFNAR2	interferon (alpha, beta and omega) receptor 2
IKBkE	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase epsilon
IKBkG	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma
ISG15	ISG15 ubiquitin-like modifier
ITCH	itchy E3 ubiquitin protein ligase
KAT2A	K(lysine) acetyltransferase 2A
KAT2B	K(lysine) acetyltransferase 2B
KCTD10	potassium channel tetramerisation domain containing 10
KCTD13	potassium channel tetramerisation domain containing 13
KIAA1530	UV-stimulated scaffold protein A
KIF23	kinesin family member 23
KIR2DL3	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 3
KLH13	Kelch-Like Family Member 13
LCK	lymphocyte-specific protein tyrosine kinase
LSM2	LSM2 homolog, U6 small nuclear RNA associated
Lys6-D	Lymphocyte antigen 6 complex locus protein G6d Precursor
MARK1	MAP/microtubule affinity-regulating kinase 1
MARK2	MAP/microtubule affinity-regulating kinase 2
MARK3	MAP/microtubule affinity-regulating kinase 3
MARK4	MAP/microtubule affinity-regulating kinase 4
MCL1	myeloid cell leukemia sequence 1
MDC1	mediator of DNA-damage checkpoint 1
MDM2	p53 E3 ubiquitin protein ligase
MDM4	p53 binding protein
MLLT4	myeloid/lymphoid or mixed-lineage leukemia translocated to 4
MTOR	mechanistic target of rapamycin
MYBPC1	myosin binding protein C
MYC	v-myc myelocytomatosis viral oncogene homolog
NUAK1	NUAK family, SNF1-like kinase, 1
OPTN	optineurin
OS9	osteosarcoma amplified 9, endoplasmic reticulum lectin
OTUB1	OTU domain, ubiquitin aldehyde binding 1
PAN3	PAN3 poly(A) specific ribonuclease subunit homolog
PHLPP1	PH domain and leucine rich repeat protein phosphatase 1
PLK1	polo-like kinase 1
PPMIG	protein phosphatase 1G
PPT1	palmitoyl-protein thioesterase 1
PRKCI	protein kinase C, iota
PRPF3	PRP3premRNA processing factor3
PSMA	proteasome (prosome, macropain) subunit, alpha
PSMB	proteasome (prosome, macropain) subunit, beta
PSMD7	proteasome (prosome, macropain) 26S subunit, non-ATPase, 7
RB1	retinoblastoma 1
RB2	retinoblastoma-like 2
RING1	ring finger protein 1
RIPK	receptor (TNFRSF)-interacting serine-threonine kinase 1
RNF2	ring finger protein 2
RNF41	ring finger protein 41
RNF128	ring finger protein 128

RNF220	ring finger protein 220
ROBO1	roundabout, axon guidance receptor, homolog 1
SAP130	Sin3A-associated protein
SART3	squamous cell carcinoma antigen recognized by T cells 3
SARS2	seryl-tRNA synthetase 2
SCF	Skp, Cullin, F-box containing complex
SERPINCI	serpin peptidase inhibitor
Sin3A	SIN3 transcription regulator homolog A
SLC25A4	solute carrier family 25 member 4
SMAD7	SMAD family member 7
SNX3	sorting nexin 3
SQSTM1	sequestosome 1
SRF	serum response factor
SUDS3	suppressor of defective silencing 3 homolog
SUMO2	SMT3 suppressor of mif two 3 homolog 2
SUMO3	SMT3 suppressor of mif two 3 homolog 3
STAGA	SPT3-TAF9-GCN5 acetylase complex
STAMs	signal transducing adaptor molecules
TBK1	TANK-binding kinase 1
TCEAL1	transcription elongation factor A (SII)-like 1
TGFβ1	Tumour beta factor beta 1
TGFβR1	Tumour beta factor beta receptor 1
TGFβR2	Tumour beta factor beta receptor 2
TMEM49	vacuole membrane protein 1
TP53BP1	tumor protein p53 binding protein1
TRAF2	TNF receptor-associated factor 2
TRAF6	TNF receptor-associated factor 6
TRAIP	TRAF interacting protein
TRRAP	transformation/transcription domain-associated protein
TSPYL4	TSPY-like 4
UBA52	ubiquitin A-52 residue ribosomal protein fusion product 1
UBB	ubiquitin B
UBC	ubiquitin C
UCHL5	ubiquitin carboxyl-terminal hydrolase L5
UFDIL	ubiquitin fusion degradation 1 like
UHRF1	ubiquitin-like with PHD and ring finger domains protein
VHL	von Hippel-Lindau tumor suppressor
WDR20	WD repeat domain 20
WDR48	WD repeat domain 48
WRNIP1	Werner helicase interacting protein 1
ZBTB32	ubiquitin-like with PHD and ring finger domains

Appendix VI. Accession number of Chitinase and Chitinase Like Protein

S.No	Species	Annotation	Nucleotide	Protein	Rec. Annotation
1	Ailuropoda melanoleuca	Chit1	XM_002925435.1	XP_002925481.1	Chit1
2	Ailuropoda melanoleuca	Chia	XM_002927484.1	XP_002927530.1	Chia1
3	Ailuropoda melanoleuca	Ovgp1	XM_002927484.1	XP_002927530.1	Ovgp1
4	Ailuropoda melanoleuca	Chi3l1	XM_002925445	XP_002925491.1	Chil1
5	Ailuropoda melanoleuca	Ctbs	XP_002930839.1	XM_002930793.1	Ctbs
6	Ailuropoda melanoleuca	Chid1	XP_002930842.1	XP_002930842	Chid1
7	Anolis carolinensis	Novel gene	ENSACAT00000025971	ENSACAP00000020256	chit1
8	Anolis carolinensis	chia	XM_003220323.1	XP_003220371.1	chia1
9	Anolis carolinensis	chia like	XM_003220322.1	XP_003220371.1	chia1
10	Anolis carolinensis	chi3l1/2	XM_003220374.1	XP_003220422.1	chil2
11	Anolis carolinensis	chia like	XM_003220321.1	XP_003220369.1	chia3
12	Anolis carolinensis	chia like	XM_003220376.1	XP_003220424.1	chio
13	Anolis carolinensis	chia like	XM_003220324.1	XP_003220372.1	chio
14	Anolis carolinensis	chia like	XM_003220526	XP_003220574.1	chio
15	Anolis carolinensis	ctbs	XM_003223062.1	XP_003223110.1	ctbs
16	Anolis carolinensis	chid1	XM_003214770.1	XP_003214818.1	chid1
17	Bos taurus	Chia	NM_174699.2	NP_777124.1	Chia1
18	Bos taurus	Ovgp1	NM_001080216.1	NP_001073685.1	Ovgp1a
19	Bos taurus	Ovgp1	XM_003585814.1	XP_003585862.1	Ovgp1b
20	Bos taurus	none	ENSBTAT00000057237	ENSBTAT00000032129	Chio1
21	Bos taurus	Chi3l1	NM_001080219.1	NP_001073688.1	Chil1
22	Bos taurus	Chi3l2	XM_003581958.1	XP_003582006.1	Chil2
23	Bos taurus	Ctbs	NM_001206600.1	NP_001193529.1	Ctbs
24	Bos taurus	Chid1	NM_001015515.1	NP_001015515.1	Chid1
25	Branchiostoma floridae	Hypothetical protein	XM_002597546.1	XP_002597592.1	cht-a
26	Branchiostoma floridae	Hypothetical protein	XM_002597545.1	XP_002597591.1	cht-b
27	Bubalus bubalis	Oviductin	EU382735.1	ABY84056.1	Ovgp1

28	Bubalus bubalis	Mammary gland protein 40	AY295929.2	AAP42568.2	BP40/Chil1
28	Bufo japonicus	Chitinase	AJ345054.1	CAC87888.1	chio2
29	Caenorabhditis elegans	Chitinase	NM_064259.4	NP_496660.2	chit
30	Callithrix jacchus	CHIT1	XM_002760664.1	XP_002760710.1	CHIT1
31	Callithrix jacchus	CHIA	XM_002751246.1	XP_002751292.1	CHIA1
32	Callithrix jacchus	OVGP1	XM_002751247.1	XP_002751293.1	OVGP1
33	Callithrix jacchus	CHI3L1	XM_002760663.1	XP_002760709.1	CHIL1
34	Callithrix jacchus	CHI3L2	XM_002751215.1	XP_002751261.1	CHIL2
35	Callithrix jacchus	CTBS	XM_002751015.1	XM_002751015.1	CTBS
36	Callithrix jacchus	CHID1	XR_089464.1	In silico translation	CHID1
37	Callithrix jacchus	Chia like (LOC100390524)	XM_002751216.1	XP_002751262.1	CHIA2
38	Canis lupus familiaris	Chia	XM_537030.3	XP_537030.3	Chia1
39	Canis lupus familiaris	Ovgp1	XM_847145.2	XP_852238.2	Ovgp1
40	Canis lupus familiaris	Chi3l1	NM_001177807.1	NP_001171278.1	Chil1
41	Canis lupus familiaris	Ctbs	XM_547309.2	XP_547309.2	Ctbs
42	Canis lupus familiaris	Chid1	XM_003432455.1	XP_003432503.1	Chid1
43	Capra hircus	Oviductin	DQ482670.1	ABF20534.1	Ovgp1
44	Capra hircus	BP40	AY081150.1	AAL87007.1	BP40/Chil1
45	Cavia porcellus	Chit1	XM_003474681	XP_003474729.1	Chit1
46	Cavia porcellus	Ovgp1	XM_003479002.1	XP_003479050.1	Ovgp1
47	Cavia porcellus	Chi3l1	XM_003474682.1	XP_003474730.1	Chil1
48	Cavia porcellus	Chid1	XM_003461302.1	XP_003461350.1	Chid1
49	Choloepus hoffmanni	Chit1	ENSCHOT00000007285	ENSCHOP00000006437	Chit1
50	Choloepus hoffmanni	Chia	ENSCHOT00000006163	ENSCHOP00000005435	Chia1
51	Choloepus hoffmanni	Ovgp1	ENSCHOT00000011219	ENSCHOP00000009901	Ovgp1
52	Choloepus hoffmanni	Chi3l1	ENSCHOT00000004023	ENSCHOP00000003546	Chil1
53	Choloepus hoffmanni	Chi3l2	ENSCHOT00000005443	ENSCHOP00000004804	Chil2
54	Ciona intestinalis	chitinase	NM_001114627.1	NP_001108099.1	cht
55	Cricetulus griseus	Chit1	XM_003498841.1	XP_003498889.1	Chit1
56	Cricetulus griseus	Chia	XM_003514388.1	XP_003514436.1	Chia1
57	Cricetulus griseus	Ovgp1	XM_003514351.1	XP_003514399.1	Ovgp1
58	Cricetulus griseus	Chi3l1	XM_003498875.1	XP_003498923.1	Chil1

59	Cricetulus griseus	Chi3l4	XM_003514389.1	XP_003514437.1	Chil3
60	Cricetulus griseus	Chid1	XM_003509772.1	XP_003509820.1	Chid1
61	Danio rerio	chia1	NM_213050.1	NP_998215.1	chit1
62	Danio rerio	chia2	NM_213249.1	NP_998414.1	chioIa
63	Danio rerio	chia3	NM_213213.1	NP_998378.1	chioIb
64	Danio rerio	chia4	NM_200446.1	NP_956740.1	chioIIa
65	Danio rerio	zgc:173927	NM001110041	NP_001103511.1	chioIIb
66	Danio rerio	CU571319.1	ENSDART00000111829	ENSDARP00000102083	chio/ovgp
67	Danio rerio	ctbs	BC095064.1	AAH95064.1	ctbs
68	Danio rerio	chid1	NM_200057.1	NP_956351.1	chid1
69	Dasypus novemcinctus	Chit1	ENSDNOT00000017334	ENSDNOP00000013434	Chit1
70	Dasypus novemcinctus	Chia	ENSDNOT00000004032	ENSDNOP00000003097	Chia1
71	Dasypus novemcinctus	Ovgp1	ENSDNOT00000017564	ENSDNOP0000001362	Ovgp1
72	Dasypus novemcinctus	Chi3l1	ENSDNOT00000009515	ENSDNOP00000007379	Chil1
73	Dasypus novemcinctus	Chi3l2	ENSDNOT00000016349	ENSDNOP00000012676	Chil2
74	Dipodomys ordii	Chit1	ENSDORT00000010763	ENSDORP00000010115	Chit1
75	Dipodomys ordii	Novel gene	ENSDORT00000003890	ENSDORP00000003632	Chia1
76	Dipodomys ordii	Ovgp1	ENSDORT00000013346	ENSDORP00000012547	Ovgp1
77	Dipodomys ordii	Chi3l1	ENSDORT00000002126	ENSDORP00000001991	Chil1
78	Dipodomys ordii	Novel gene	ENSDORT00000003888	ENSDORP00000003630	Chil2
79	Dipodomys ordii	Ctbs	ENSDORT00000015059	ENSDORP00000014176	Ctbs
80	Dipodomys ordii	Chid1	ENSDORT00000015451	ENSDORP00000014543	Chid1
81	Drosophila melanogaster	cht2	NM_057950.2	NP_477298.2	cht2
82	Drosophila melanogaster	cht3	NM_001042957.1	NP_001036422.1	cht3
83	Drosophila melanogaster	cht4	NM_080223.2	NP_524962.2	cht4
84	Drosophila melanogaster	cht5	NM_142057.2	NM_142057.2	cht5
85	Drosophila melanogaster	cht6	NM_132370.2	NP_572598.2	cht6
86	Drosophila melanogaster	cht7	NM_139511.3	NP_647768.3	cht7
87	Drosophila melanogaster	cht8	NM_137698.1	NP_611542.2	cht8
88	Drosophila melanogaster	cht9	NM_137699.4	NP_611543.3	cht9
89	Drosophila melanogaster	cht11	NM_132133.2	NP_572361.1	cht11
90	Drosophila melanogaster	cht12	NM_166420.2	NP_726022.1	cht12
91	Drosophila melanogaster	IDGF1	AF102236.1	AAC99417.1	IDGF1
92	Drosophila melanogaster	IDGF2	AF102237.1	AAC99418.1	IDGF2

93	Drosophila melanogaster	IDGF3	AF102238.1	AAC99419.1	IDGF3
94	Drosophila melanogaster	IDGF4	AF102239	AAC99420.1	IDGF4
95	Drosophila melanogaster	IDGF5	NM_137477.3	NP_611321.3	IDGF5
96	Drosophila melanogaster	GH20192	BT029919.1	ABM92793.1	IDGF6
97	Drosophila melanogaster	CG8460	NM_135346.3	NP_609190.2	--
98	Echinops telfairi	Chit1	ENSETET00000001935	ENSETEP00000001566	Chit1
99	Echinops telfairi	Chia	ENSETET00000011092	ENSETEP00000008995	Chia1
100	Echinops telfairi	Ovgp1	ENSETET00000002991	ENSETEP00000002454	Ovgp1
101	Echinops telfairi	Chi3l1	ENSETET00000013193	ENSETEP00000010701	Chil1
102	Echinops telfairi	Chi3l2	ENSETET00000004415	ENSETEP00000003616	Chil2
103	Echinops telfairi	Chid1	ENSETET00000014627	ENSETEP00000011857	Chid1
104	Equus caballus	Chit1	NM_001143797.1	NP_001137269.1	Chit1
105	Equus caballus	Ovgp1	ENSECAT00000024758	ENSECAP00000020584	Ovgp1
106	Equus caballus	Chi3l1	XM_001496500.2	XP_001496550.2	Chil1
107	Equus caballus	None	ENSECAT00000025588	ENSECAP00000021290	Chil2
108	Equus caballus	Ctbs	XM_003365103.1	XP_003365151.1	Ctbs
109	Equus caballus	Chid1	XM_003362643.1	XP_003362691.1	Chid1
110	Erinaceus europaeus	Chia	ENSEEUT00000005347	ENSEEUP00000004866	Chia1
111	Erinaceus europaeus	Ovgp1	ENSEEUT00000011672	ENSEEUP00000010650	Ovgp1
112	Erinaceus europaeus	Chi3l1	ENSEEUT00000011096	ENSEEUP00000010116	Chil1
113	Felis catus	Chit1	ENSFCAT00000013116	ENSFCAP00000012160	Chit1
114	Felis catus	Chi3l1	ENSFCAT00000013115	ENSFCAP00000012159	Chil1
115	Felis catus	Ctbs	ENSFCAT00000008799	ENSFCAP00000008157	Ctbs
116	Felis catus	Chid1	ENSFCAT00000004824	ENSFCAP00000004457	Chid1
117	Gadus morhua	chia1	ENSGMOT00000016457	ENSGMOP00000016048	PS
118	Gadus morhua	chia2	ENSGMOT00000016435	ENSGMOP00000016026	“
119	Gadus morhua	None	ENSGMOT00000021699	ENSGMOP00000021181	“
120	Gadus morhua	ovgp1(1)	ENSGMOT00000017220	ENSGMOP00000016801	“
121	Gadus morhua	ovgp1(2)	ENSGMOT00000011353	ENSGMOP00000011053	“
122	Gadus morhua	ovgp1(3)	ENSGMOT00000017190	ENSGMOP00000016771	“
123	Gadus morhua	ovgp1(4)	ENSGMOT00000011369	ENSGMOP00000011068	“
124	Gadus morhua	ctbs	ENSGMOT00000017946	ENSGMOP00000017514	ctbs
125	Gadus morhua	chid1	ENSGMOT00000019384	ENSGMOP00000018925	chid1
126	Gallus gallus	E1BZP6	ENSGALT00000005566	ENSGALP00000005553	chia1a

		(chia)			
127	Gallus gallus	E1BZP4 (chia)	ENSGALT00000005565	ENSGALP00000005555	chia1b
128	Gallus gallus	F1NMM2	ENSGALT00000005564	ENSGALP00000005554	chia1c
129	Gallus gallus	ctbs	XM_422372.3	XP_422372.1	ctbs
130	Gallus gallus	chid1	NM_001199634.1	NP_001186563.1	chid1
131	Gasterosteus aculeatus	chia	ENSGACT00000000527	ENSGACP00000000527	chia1a
132	Gasterosteus aculeatus	chia	ENSGACT00000016542	ENSGACP00000016509	chia1b
133	Gasterosteus aculeatus	Novel gene	ENSGACT00000015229	ENSGACP00000015200	chio/ovgp
134	Gasterosteus aculeatus	Novel gene	ENSGACT00000004677	ENSGACP00000004653	chioIa
135	Gasterosteus aculeatus	Novel gene	ENSGACT00000004692	ENSGACP00000004687	chioIb
136	Gasterosteus aculeatus	Novel gene	ENSGACT00000015218	ENSGACP00000015189	Chit1
137	Gasterosteus aculeatus	Novel gene	ENSGACT00000016635	ENSGACP00000016602	chioII/chil
138	Gasterosteus aculeatus	ctbs	ENSGACT00000012822	ENSGACP00000012798	ctbs
139	Gasterosteus aculeatus	chid1	ENSGACT00000007086	ENSGACP00000007068	chid1
140	Gorilla gorilla	CHIT	ENSGGOT00000011232	ENSGGOP00000010909	CHIT
141	Gorilla gorilla	CHIA	ENSGGOT00000008564	ENSGGOP00000008334	CHIA1
142	Gorilla gorilla	OVGP1	ENSGGOT00000005228	ENSGGOP00000005097	OVGP 1
143	Gorilla gorilla	CHI3L1	ENSGGOT00000009955	ENSGGOP00000009685	CHIL1
144	Gorilla gorilla	CHI3L2	ENSGGOT00000016075	ENSGGOP00000015627	CHIL2
145	Gorilla gorilla	CTBS	ENSGGOT00000015792	ENSGGOP00000015353	CTBS
146	Gorilla gorilla	CHID1	ENSGGOT00000010884	ENSGGOP00000010570	CHID1
147	Homo sapiens	CHIT1	BC105680.1	AAI05681.1	Chit1
148	Homo sapiens	CHIA	AF290004.1	AAG60019.1	Chia1
149	Homo sapiens	OVGP1	BC126177.1	AAI26178.1	Ovgp1
150	Homo sapiens	CHI3L1	NM_001276.2	NP_001267.2	CHIL1
151	Homo sapiens	CHI3L2	NM_001025197.1	NP_001020368.1	CHIL2
152	Homo sapiens	CTBS	BC126333.1	AAI26334.1	CTBS
153	Homo sapiens	CHID1	NM_001142675.1	NP_001136147.1	CHID1
154	Homo sapiens	CHIA-pseudo	RP11-165H20.1	--	CHIA2-pseudo
155	Homo sapiens	CHIA-pseudo	RP11-165H20.4	--	CHIA3-pseudo
156	Homo sapiens	CHIA-pseudo	RP11-1125M8.5	--	CHIO-pseudo
157	Lethenteron japonicum	chit	EU741679.1	ACF10400.1	cht
158	Loxodonta africana	Chit1	XM_003410110.1	XP_003410158.1	Chit1
159	Loxodonta africana	Chia	XM_003409404.1	XP_003409452.1	Chia1
160	Loxodonta africana	Ovgp1	XM_003409406.1	XP_003409454.1	Ovgp1
161	Loxodonta africana	Chi3l1	XM_003410238.1	XP_003410286.1	Chil1
162	Loxodonta africana	Chi3l2	XM_003409572.1	XP_003409620.1	Chil2
163	Loxodonta africana	Ctbs	ENSLAFT00000011740	ENSLAFP00000009812	Ctbs
164	Loxodonta africana	Chid1	XM_003423324.1	XP_003423372.1	Chid1
165	Macaca mulatta	CHIT1	XM_001103012.2	XP_001103012.1	CHIT1

166	Macaca mulatta	CHIA	ENSMMUT00000012389	ENSMMUP00000011619	CHIA1
167	Macaca mulatta	OVGP1	NM_001042787.1	NP_001036252.1	OVGP1
168	Macaca mulatta	CHI3L1	XM_001103739.2	XP_001103739.1	CHIL1
169	Macaca mulatta	CHI3L2	XM_001093397.2	XP_001093397.2	CHIL2
170	Macaca mulatta	CTBS	XM_001107057.2	XP_001107057.1	CTBS
171	Macaca mulatta	CHID1	XM_001089724.2	XP_001089724.1	CHID1
172	Macaca mulatta	pseudo	LOC705382	--	CHIA2-pseudo
173	Macaca mulatta	pseudo	LOC100425748	--	CHIA3-pseudo
174	Macaca mulatta	pseudo	LOC100425497	--	CHIO-pseudo
175	Macropus eugenii	Chit1	ENSMEUT00000001040	ENSMEUP00000000957	Chit1
176	Macropus eugenii	Chia	ENSMEUT00000010728	ENSMEUP00000009760	Chia1
177	Macropus eugenii	Chi3l1	ENSMEUT00000009207	ENSMEUP00000008387	Chil1
178	Macropus eugenii	Chi3l2	ENSMEUT00000011803	ENSMEUP00000010725	Chil2
179	Macropus eugenii	Ctbs	ENSMEUT00000007763	ENSMEUP00000007067	Ctbs
180	Macropus eugenii	Chid1	ENSMEUT00000006175	ENSMEUP00000005628	Chid1
181	Meleagris gallopavo	chia	XM_003212987.1	XP_003213035.1	chia1a
182	Meleagris gallopavo	chia	XM_003212986.1	XP_003213034.1	chia1c
183	Meleagris gallopavo	chia	XM_003212985.1	XP_003213033.1	chia1b
184	Meleagris gallopavo	ctbs	XM_003208743.1	XP_003208791.1	ctbs
185	Meleagris gallopavo	chid1	XM_003206307.1	XP_003206355.1	chid1
186	Mesocricetus auratus	Ovgp1	D32218.1	BAA06977.1	Ovgp1
187	Microcebus murinus	CHIT1	ENSMICT00000002309	ENSMICP00000002109	CHIT1
188	Microcebus murinus	CHIA	ENSMICT00000011956	ENSMICP00000010886	CHIA1
189	Microcebus murinus	CHI3L1	ENSMICT00000015438	ENSMICP00000014068	CHIL1
190	Microcebus murinus	CHI3L2	ENSMICT00000011945	ENSMICP00000010875	CHIL2
191	Microcebus murinus	CTBS	ENSMICT00000014904	ENSMICP00000013589	CTBS
192	Microcebus murinus	CHID1	ENSMICT00000000972	ENSMICP00000000880	CHID1
193	Monodelphis domestica	Chit1	XM_001369883.2	XP_001369920.2	Chit1
194	Monodelphis domestica	Chia	XM_001372827.1	XP_001372864.1	Chia1a
195	Monodelphis domestica	Chia like	XM_001372844.1	XP_001372881.1	Chia1b
196	Monodelphis domestica	Chia like	XM001381953.1	XP_001381990.1	Chia3
197	Monodelphis domestica	Chia like	XM_001381960.2	XP_001381997.2	Chio1a

198	Monodelphis domestica	Chia like	XM_001381962.2	XP_001381999.2	Chio1b
199	Monodelphis domestica	Ovgp1	ENSMODT00000001650	ENSMODP00000001616	Ovgp1
200	Monodelphis domestica	Chi3l1	XM_001364797.2	XP_001364834.2	Chil1
201	Monodelphis domestica	Chi3l2	XM_001381951.2	XP_001381988.2	Chil2
202	Monodelphis domestica	Ctbs	XM_001367032.1	XP_001367069.1	Ctbs
203	Monodelphis domestica	Chid1	XM_003342209.1	XP_003342257.1	Chid1
204	Mus musculus	Chit1	BC138765.1	AAI38766.1	Chit1
205	Mus musculus	Chia	DQ349202.1	ABC86699.1	Chia
206	Mus musculus	Ovgp1	BC137995	AAI37996.1	Ovgp1
207	Mus musculus	Chi3l1	NM_007695.3	NP_031721.2	Chil1
208	Mus musculus	Chi3l3	NM_009892.2	NP_034022.2	Chil3
209	Mus musculus	Chi3l4	NM_145126.2	NP_660108.2	Chil4
210	Mus musculus	Chi3l7	XM_003086317.1	XP_003086365.1	Chil7
211	Mus musculus	BYm	NM_178412.2	NP_848499.1	Chil8
212	Mus musculus	Ctbs	NM_028836.3	NP_083112.1	Ctbs
213	Mus musculus	Chid1	BC061063.1	AAH61063.1	Chid1
214	Myotis lucifugus	novel gene	ENSMLUT00000010603	ENSMLUP00000009664	Chia1a
215	Myotis lucifugus	novel gene	ENSMLUT00000008998	ENSMLUP00000008199	Chia1b
216	Myotis lucifugus	Ovgp1	ENSMLUT00000012410	ENSMLUP00000011299	Ovgp1
217	Myotis lucifugus	novel gene	ENSMLUT00000023766	ENSMLUP00000017407	Chil1
218	Myotis lucifugus	novel gene	ENSMLUT00000010322	ENSMLUP00000009406	Chil1
219	Myotis lucifugus	Ctbs	ENSMLUT00000016040	ENSMLUP00000014612	Ctbs
220	Myotis lucifugus	Chid1	ENSMLUT00000016623	ENSMLUP00000015148	Chid1
221	Nomascus leucogenys	CHIT1	XM_003264580.1	XP_003264628.1	CHIT1
222	Nomascus leucogenys	CHIA	XM_003267959.1	XP_003268007.1	CHIA1
223	Nomascus leucogenys	OVGP1	XM_003267966.1	XP_003268014.1	OVGP1
224	Nomascus leucogenys	CHI3L1	XM_003264578.1	XP_003264626.1	CHIL1
225	Nomascus leucogenys	CHI3L2	XM_003267958.1	XP_003268006.1	CHIL2
226	Nomascus leucogenys	CTBS	XM_003278435.1	XP_003278483.1	CTBS
227	Nomascus leucogenys	CHID1	XM_003281317.1	XP_003281365.1	CHID1
228	Ochotona princeps	Chit1	ENSOPRT00000014806	ENSOPRP00000013520	Chit1
229	Ochotona princeps	Ovgp1	ENSOPRT00000003825	ENSOPRP00000003520	Ovgp1

230	Ochotona princeps	Chi3l1	ENSOPRT00000014785	ENSOPRP00000013500	Chil1
231	Ochotona princeps	Chi3l2	ENSOPRT00000009439	ENSOPRP00000008637	Chil2
232	Ochotona princeps	Chid1	ENSOPRT00000001568	ENSOPRP00000001445	Chid1
233	Oncorhynchus mykiss	Gastric chitinase	EU877960.1	ACG58867.1	Chia1
234	Oncorhynchus mykiss	chit	AJ535688.1	CAD59687.1	chioII
235	Oreochromis niloticus	chialike	XM_003459038.1	XP_003438829.1	chit1a
236	Oreochromis niloticus	chialike	XM_003459039.1	XP_003459087.1	chit1b
237	Oreochromis niloticus	chialike	XM_003458530.1	XP_003458578.1	chit1c
238	Oreochromis niloticus	chialike	XM_003438781.1	XP_003438829.1	chioI
239	Oreochromis niloticus	ctbs	XM_003452536.1	XP_003452584.1	ctbs
240	Oreochromis niloticus	chid1	XM_003440467.1	XP_003440515.1	chid1
241	Ornithorhynchus anatinus	Chit1	XM_001518544.2	XP_001518594.2	Chit1
242	Ornithorhynchus anatinus	novel	ENSOANT00000015751	ENSOANP00000015748	Chit2
243	Ornithorhynchus anatinus	novel	ENSOANT00000003822	ENSOANP00000003821	Chit3
244	Ornithorhynchus anatinus	Ovgp1	XM_001517948.2	XP_001517998.2	Ovgp1
245	Ornithorhynchus anatinus	Chi3l1	XM_001518538.2	XP_001518588.2	Chil1
246	Ornithorhynchus anatinus	Ctbs	XM_001514909.1	XP_001514959.1	Ctbs
247	Ornithorhynchus anatinus	Chid1	XM_001515150.2	XP_001515200.2	Chid1
248	Oryctolagus cuniculus	Chit1	XM_002717457.1	XP_002717503.1	Chit1
249	Oryctolagus cuniculus	Ovgp1	NM_001082105.1	NP_001075574.1	Ovgp1
250	Oryctolagus cuniculus	Chi3l1	XM_002717458.1	XP_002717504.1	Chil1
251	Oryctolagus cuniculus	Chi3l2	XM_002715779.1	XP_002715825.1	Chil2
252	Oryctolagus cuniculus	Ctbs	XM_002715538.1	XP_002715584.1	Ctbs
253	Oryctolagus cuniculus	Chid1	XM_002724161.1	XP_002724207.1	Chid1
254	Oryzias latipes	Novel gene	ENSORLT00000013331	ENSORLP00000013330	chioI
255	Oryzias latipes	Novel gene	ENSORLT00000017136	ENSORLP00000017135	chioIIa
256	Oryzias latipes	Novel gene	ENSORLT00000017096	ENSORLP00000017095	chioIIb
257	Oryzias latipes	Novel gene	ENSORLT00000013269	ENSORLP00000013268	chit1
258	Oryzias latipes	Novel gene	ENSORLT00000013259	ENSORLP00000013258	chit2
259	Oryzias latipes	ctbs	ENSORLT00000007814	ENSORLP00000007813	ctbs
260	Oryzias latipes	chid1	ENSORLT00000008210	ENSORLP00000008209	chid1
261	Otolemur	CHIT1	ENSOGAT00000000928	ENSOGAP00000000830	CHIT1

	garnettii				
262	Otolemur garnettii	CHIA	ENSOGAT00000010811	ENSOGAP00000009675	CHIA1
263	Otolemur garnettii	novel gene	ENSOGAT00000010810	ENSOGAP00000009674	CHIA2
264	Otolemur garnettii	novel gene	ENSOGAT00000033204	ENSOGAP00000021751	CHIA3
265	Otolemur garnettii	CHI3L1	ENSOGAT00000000924	ENSOGAP00000000826	CHIL1
266	Otolemur garnettii	CHI3L2	ENSOGAT00000010808	ENSOGAP00000009672	CHIL2
267	Otolemur garnettii	CTBS	ENSOGAT00000004652	ENSOGAP00000004154	CTBS
268	Otolemur garnettii	CHID1	ENSOGAT00000031647	ENSOGAP00000021384	CHID1
269	Ovis aries	Chia	EF063144.1	ABP98946	Chia1
270	Ovis aries	Ovgp1	NM_001009779.1	NP_001009779.1	Ovgp1
271	Ovis aries	BP40	AY392761.1	AAQ94054.1	Chil1/BP40
272	Ovis aries	Chid1	EF581383.1	ABQ51216.1	Chid1
273	Pan troglodytes	CHIT1	XM_514112.3	XP_514112.3	CHIT1
274	Pan troglodytes	CHIA	XR_024811.2	In silico translation	CHIA1
275	Pan troglodytes	OVGP1	XM_003338996.1	XP_003339044.1	OVGP1
276	Pan troglodytes	CHI3L1	XM_001153636.2	XP_001153636.2	CHIL1
277	Pan troglodytes	CHI3L2	XM_513645.3	XP_513645.3	CHIL2
278	Pan troglodytes	CTBS	XM_513520.3	XP_513520.1	CTBS
279	Pan troglodytes	CHID1	XM_001151962.2	XP_001151962.2	CHID1
280	Papio anubis	OVGP1	NM_001112617.1	NP_001106087.1	OVGP1
281	Pelodiscus sinensis	chia	ENSPSIT00000012771	ENSPSIP00000012710	chia
282	Pelodiscus sinensis	ovgp1	ENSPSIT00000011841	ENSPSIP00000011784	chio
283	Pelodiscus sinensis	Novel gene	ENSPSIT00000011195	ENSPSIP00000011139	chio
284	Pelodiscus sinensis	Novel gene	ENSPSIT00000014733	ENSPSIP00000014664	chio
285	Pelodiscus sinensis	ctbs	ENSPSIT00000019583	ENSPSIP00000019492	ctbs
286	Pelodiscus sinensis	chid1	ENSPSIT00000015979	ENSPSIP00000015904	chid1
287	Petromyzon marinus	chid1	ENSPMAT00000009060	ENSPMAP00000009021	chid1
288	Pongo abelii	CHIT1	XM_002809597.1	XP_002809643.1	CHIT1
289	Pongo abelii	CHIA	XM_002810445.1	XP_002810491.1	CHIA1
290	Pongo abelii	OVGP1	XM_002810442.1	XP_002810488.1	OVGP1
291	Pongo abelii	CHI3L1	NM_001131991.1	NP_001125463.1	CHIL1
292	Pongo abelii	CHI3L2	ENSPPYT00000001247	ENSPPYP00000001207	CHIL2
293	Pongo abelii	CTBS	ENSPPYT00000001438	ENSPPYP00000001393	CTBS
294	Pongo abelii	CHID1	NM_001133685.1	NP_001127157.1	CHID1
295	Procavia capensis	Chit1	ENSPCAT00000003734	ENSPCAP00000003506	not assigned
296	Procavia capensis	Ovgp1	ENSPCAT00000003785	ENSPCAP00000003553	“
297	Procavia capensis	Chi3l1	ENSPCAT00000010583	ENSPCAP00000009872	“
298	Procavia	Chi3l2	ENSPCAT00000000396	ENSPCAP00000000369	“

	capensis				
299	Procavia capensis	Chid1	ENSPCAT00000010327	ENSPCAP00000009637	“
300	Pteropus vampyrus	Ovgp1	ENSPVAT00000010601	ENSPVAP00000009989	Ovgp1
301	Pteropus vampyrus	Ctbs	ENSPVAT00000008138	ENSPVAP00000007688	Ctbs
302	Pteropus vampyrus	Chid1	ENSPVAT00000001560	ENSPVAP00000001470	Chid1
303	Rana catesbeiana	chit	AF447579.1	AAL38179.1	chia1
304	Rattus norvegicus	Chit1	NM_001079689.1	NP_001073157.1	Chit1
305	Rattus norvegicus	Chia	NM_207586.1	NP_997469.1	Chia1
306	Rattus norvegicus	Chi3l1	NM_053560.1	NP_446012.1	Chil1
307	Rattus norvegicus	Chi3l3	NM_001191712.1	NP_001178641.1	Chil3
308	Rattus norvegicus	Chi3l4	XM_227566.5	XP_227566.4	Chil7
309	Rattus norvegicus	LOC295352	NM_001134512.1	NP_001127984.1	Chil8
310	Rattus norvegicus	Ctbs	NM_031023.1	NP_112285.1	Ctbs
311	Rattus norvegicus	Chid1	NM_001047854.2	NP_001041319.2	Chid1
312	Sarcophilus harrisii	Chit1	ENSSHAT00000019889	ENSSHAP00000019732	Chit1
313	Sarcophilus harrisii	novel	ENSSHAT00000014785	ENSSHAP00000014661	Chia1a
314	Sarcophilus harrisii	Novel	ENSSHAT00000014578	ENSSHAP00000014457	Chia3
315	Sarcophilus harrisii	Novel	ENSSHAT00000014916	ENSSHAP00000014791	Chia1b
316	Sarcophilus harrisii	novel	ENSSHAT00000015082	ENSSHAP00000014957	Chio1
317	Sarcophilus harrisii	Ovgp1	ENSSHAT00000015439	ENSSHAP00000015312	Ovgp1
318	Sarcophilus harrisii	Chi3l1	ENSSHAT00000019702	ENSSHAP00000019545	Chil1
319	Sarcophilus harrisii	Chi3l2	ENSSHAT00000014471	ENSSHAP00000014352	Chil2
320	Sarcophilus harrisii	Ctbs	ENSSHAT00000020721	ENSSHAP00000020557	Ctbs
321	Sarcophilus harrisii	Chid1	ENSSHAT00000003724	ENSSHAP00000003687	Chid1
322	Sorex araneus	Chia	ENSSART00000005351	ENSSARP00000004846	Chia
323	Sorex araneus	Ovgp1	ENSSART00000002045	ENSSARP00000001859	Ovgp1
324	Sorex araneus	Ctbs	ENSSART00000002000	ENSSARP00000001819	Ctbs
325	Sorex araneus	Chid1	ENSSART00000008518	ENSSARP00000007704	Chid1
326	Spermophilus tridecemlineatus	Chit1	ENSSTOT00000023652	ENSSTOP00000013500	Chit1
327	Spermophilus tridecemlineatus	Chia	ENSSTOT00000002559	ENSSTOP00000002289	Chia1a
328	Spermophilus tridecemlineatus	Chia	ENSSTOT00000005561	ENSSTOP00000004975	Chia1b
329	Spermophilus tridecemlineatus	Ovgp1	ENSSTOT00000000474	ENSSTOP00000000426	Ovgp1
330	Spermophilus tridecemlineatus	Chi3l1	ENSSTOT00000015439	ENSSTOP00000013826	Chil1

331	Spermophilus tridecemlineatus	Ctbs	ENSSTOT00000015317	ENSSTOP00000013724	Ctbs
332	Spermophilus tridecemlineatus	Chid1	ENSSTOT00000006593	ENSSTOP00000005894	Chid1
333	Sus scrofa	Chit1	XM_003130296.1	XP_003130344.1	Chit1
334	Sus scrofa	Chia	XR_130567.1	translation	Chia
335	Sus scrofa	Ovgp1	NM_214070.1	NP_999235.1	Ovgp1
336	Sus scrofa	Heparin binding protein	U19900.1	AAA86482	Chil1
337	Sus scrofa	BP40	AY762599.1	AAV30548.1	BP40/Chil1
338	Sus scrofa	Chi3l2	XM_003481491.1	XP_003481539.1	Chil2
339	Sus scrofa	Ctbs	XM_003356418.2	XP_003356466.2	Ctbs
340	Sus scrofa	Chid1	NM_001243810.1	NP_001230739.1	Chid1
341	Taeniopygia guttata	chia	ENSTGUT00000018139	ENSTGUP00000017736	chia1
342	Taeniopygia guttata	ctbs	ENSTGUT00000006829	ENSTGUP00000006761	ctbs
343	Taeniopygia guttata	chid1	ENSTGUT00000010151	ENSTGUP00000010044	chid1
344	Tarsius syrichta	CHIT1	ENSTSYT00000006197	ENSTSYP00000005672	CHIT1
345	Tarsius syrichta	CHIA	ENSTSYT00000011237	ENSTSYP00000010306	CHIA1
346	Tarsius syrichta	CHI3L1	ENSTSYT00000005582	ENSTSYP00000005109	CHIL1
347	Tarsius syrichta	CHI3L2	ENSTSYT00000002608	ENSTSYP00000002398	CHIL2
348	Tarsius syrichta	CTBS	ENSTSYT00000010032	ENSTSYP00000009202	CTBS
349	Tarsius syrichta	CHID1	ENSTSYT00000009359	ENSTSYP00000008587	CHID1
350	Tetraodon nigroviridis	Novel gene	ENSTNIT00000002411	ENSTNIP00000000910	chioI
351	Tetraodon nigroviridis	Novel gene	ENSTNIT00000003839	ENSTNIP00000001773	chioII
352	Tetraodon nigroviridis	Novel gene	ENSTNIT00000002465	ENSTNIP00000003161	chit1
353	Tetraodon nigroviridis	ctbs	ENSTNIT00000018854	ENSTNIP00000018627	ctbs
354	Tetraodon nigroviridis	chid1	ENSTNIT00000019388	ENSTNIP00000019160	chid1
355	Tupaia belangeri	Chia	ENSTBET00000006966	ENSTBEP00000006015	Chia1
356	Tupaia belangeri	Ovgp1	ENSTBET00000009993	ENSTBEP00000008647	Ovgp1
357	Tupaia belangeri	Chi3l1	ENSTBET00000010066	ENSTBEP00000008711	Chil1
358	Tupaia belangeri	Chi3l2	ENSTBET00000011188	ENSTBEP00000009665	Chil2
359	Tupaia belangeri	Ctbs	ENSTBET00000006121	ENSTBEP00000005268	Ctbs
360	Tupaia belangeri	Chid1	ENSTBET00000012200	ENSTBEP00000010565	Chid1
361	Tursiops truncatus	Chit1	ENSTTRT00000000835	ENSTTRP00000000788	Chit1
362	Tursiops truncatus	Chia	ENSTTRT00000007832	ENSTTRP00000007407	Chia1
363	Tursiops truncatus	Ovgp1	ENSTTRT00000012152	ENSTTRP00000011522	Ovgp1
364	Tursiops truncatus	Chi3l1	ENSTTRT00000009513	ENSTTRP00000009014	Chil1
365	Tursiops truncatus	Chi3l2	ENSTTRT00000007808	ENSTTRP00000007384	Chil2
366	Tursiops truncatus	Ctbs	ENSTTRT00000012373	ENSTTRP00000011740	Ctbs
367	Tursiops	Chid1	ENSTTRT00000012505	ENSTTRP00000011864	Chid1

	truncatus				
368	Takifugu rubripes	Novel gene	ENSTRUT00000001850	ENSTRUP00000001842	chioIa
369	Takifugu rubripes	Novel gene	ENSTRUT00000022661	ENSTRUP00000022567	chioIb
370	Takifugu rubripes	Novel gene	ENSTRUT00000033584	ENSTRUP00000033458	chioIIa
371	Takifugu rubripes	Novel gene	ENSTRUT00000006584	ENSTRUP00000006542	chioIIb
372	Takifugu rubripes	Novel gene	ENSTRUT00000044944	ENSTRUP00000044793	ctbs
373	Takifugu rubripes	chid1	ENSTRUT00000009179	ENSTRUP00000009125	chid1
374	Vicugna pacos	Chit1	ENSVPAT00000008798	ENSVPAP00000008186	Chit1
375	Vicugna pacos	Chia	ENSVPAT00000000342	ENSVPAP00000000318	Chia1
376	Vicugna pacos	Ovgp1	ENSVPAT00000000345	ENSVPAP00000000321	Ovgp1
378	Vicugna pacos	Chi3l1	ENSVPAT00000003605	ENSVPAP00000003341	Chil1
379	Vicugna pacos	Ctbs	ENSVPAT00000000313	ENSVPAP00000000290	Ctbs
380	Xenopus laevis	chia	AF447580.1	AAL38180.1	chia
381	Xenopus laevis	chit	BC073276.1	AAH73276.1	chio
382	Xenopus laevis	ctbs	NM_001094061.1	NP_001087530.1	ctbs
383	Xenopus laevis	chid1	NM_001086262.1	NP_001079731.1	chid1
384	Xenopus tropicalis	chia	NM_001199560.1	NP_001186489.1	chia1
385	Xenopus tropicalis	XB-GENE-5763443	ENSXETT00000025880	ENSXETP00000025880	chia3
386	Xenopus tropicalis	chit1	NM_001005792.1	NP_001005792.1	chio
387	Xenopus tropicalis	Ctbs	ENSXETT00000066334	ENSXETP00000059660	ctbs
388	Xenopus tropicalis	Chid1	ENSXETT00000041328	ENSXETP00000041328	chid1