Profiling, Comparison and Validation of Gene Expression in Gastric Carcinoma and Normal Stomach

Karin A Øien

BSc, MBChB, MRCPath

A thesis submitted to the University of Glasgow
in partial fulfilment of the requirements
for the Degree of Doctor of Philosophy

April 2002

Cancer Research UK Department of Medical Oncology
and Department of Pathology,
University of Glasgow

Abstract

Gastric carcinoma is the second most common cause of cancer death world-wide but its molecular biology is not well understood. My aims were to catalogue the genes expressed in gastric carcinoma and normal stomach and to identify differentially expressed and gastric-specific transcripts. Serial analysis of gene expression (SAGE) produces comprehensive, quantitative and reproducible expression profiles. The method of SAGE was established in this laboratory then used to study normal gastric antral mucosa and two gastric adenocarcinomas of distal, intestinal type. The libraries were compared on-line with other glandular epithelial tissues. Selected genes were validated in a panel of 19 normal and tumour gastro-intestinal tissues and cell lines by Northern blotting and immunohistochemistry.

29,480 transcripts, derived from 10,866 genes, were identified. The validation studies corroborated the SAGE profiles although tumour heterogeneity was noted. 1% of genes were differentially expressed (by over five-fold and with a p-value below 0.01) between the pooled gastric carcinomas and normal stomach. The most abundant transcripts included ribosomal and mitochondrial proteins, of which most were up-regulated in the tumours, as were other widely expressed genes including transcription factors (Id1), signalling molecules (fibroblast growth factor receptor and serine/threonine protein kinases), coatomer and proteasome components, thymosin beta 10 and collagenase I. In contrast, cytoskeletal proteins (alpha actinin and profilin) were down-regulated in the tumours.

Many genes which were more highly expressed in normal stomach are important in normal gastric function, including gastrin, immunoglobulin alpha, lysozyme, mucin (MUC5), trefoil peptides (pS2 and spasmolytic polypeptide) and pepsinogens, which were amongst 55 gastric-specific transcripts. Some transcripts had previously been characterised only minimally (prostate stem cell antigen) or not at all (aquaporin 5) in the stomach. Some genes (intestinal trefoil factor) which were up-regulated in gastric carcinoma reflect the intestinal-type histology. Some genes abundant in normal gastric antrum had previously been regarded as markers of pancreatic carcinoma. Many differentially expressed species, some tumour-associated, were novel and await investigation.

One new gene which was identified was highly expressed in normal stomach but absent from gastric carcinomas. This new gene was selected for further investigation. The SAGE expression profile was confirmed by Northern blotting and *in situ* hybridisation by which the mRNA was located in the superficial/foveolar (pit) epithelium of the gastric mucosa, so the gene was termed foveolin. The transcript was expressed outwith the stomach only in metaplastic gastric epithelium, in Barrett's oesophagus or the ulcer-associated cell lineage in the gut, and outwith the gut only in ovarian mucinous tumours. The mRNA was present in the stomach of mouse, rat and cow in the same location as in humans.

The 5' and 3' ends of the mRNA were characterised by Rapid Amplification of cDNA Ends (RACE). Homologous mouse and cow mRNAs were identified, characterised and compared. The full-length genomic sequences for the human and mouse were obtained using on-line databases, characterised and compared. A partial human genomic clone was obtained from a PAC library, and used to map the gene by fluorescent *in situ* hybridisation (FISH) to human chromosome 2. The predicted protein product, like the mRNA and DNA sequences, is highly conserved between the human, mouse and cow species. The protein shows no homology to any known protein sequences or motifs, but bears an initial signal peptide and is therefore predicted to have an extracellular location, being either retained on the outer cell surface or secreted into the gastric lumen, much like gastric mucin (MUC5) and the trefoil peptide pS2 (TFF1), with which foveolin shares a similar location in the superficial and foveolar gastric epithelium.

These are the first global profiles of gene expression in the stomach. The molecular anatomy correlated with the morphology. The gastric carcinoma profiles resembled other tumours, which supports the existence of common cancer-associated molecular targets. The normal gastric profile differed from other normal glandular tissues but agreed with existing literature. Many new transcripts were identified, of which one has been further characterised here in its first detailed description. These data increase our knowledge about the genes involved in normal gastric function and in malignant change in the stomach, and provide a catalogue of candidates from which to develop markers for better diagnosis and therapy of gastric carcinoma.

Table of Contents

List of Tables

List of figures

Acknowledgements

Thanks for making this research possible, and for making it a superb and enjoyable learning experience, to:

Nicol Keith for a highly appropriate project, excellent supervision, enthusiasm throughout and most of all extreme patience (greatly appreciated!)

Donny Black for sound advice in a familiar accent

Professors Bob Brown, Stan Kaye and Jim Cassidy for enabling (and welcoming) me to work in the Cancer Research UK Department of Medical Oncology

Professors Roddy MacSween and Barry Gusterson for allowing me to work outwith the University Department of Pathology and everyone there for their encouragement especially Ian Brown, Alan Foulis, James Going, George Lindop, Anne Marie McNicol and Robin Reid

Professors Kenneth McColl and Emad El-Omar for educating and interesting me in the stomach, *Helicobacter pylori* and gastric carcinoma

Thanks for scientific and technical help (in temporal order) to:

Stacey Hoare for help with getting started in a molecular biology laboratory in the first place and with cell culture

Grant Fullarton for surgical skills

Keith Vass for help with computing and bioinformatics, especially setting up the SAGE program and databases locally, using the UNIX servers and sequence formatting

Robert McFarlane for large-scale sequencing

Lisa Bryce for advice on RNA work, including Northern blotting and RT-PCR, and for help with analysing cDNA sequences and designing oligonucleotide primers

Sharon Burns for help with investigating selected candidate genes by TOPO cloning, plasmid preparations, sequencing and fluorescent *in situ* hybridisation

Ian Downie for immunohistochemistry

Rod Ferrier and Sharlene Butler for *in situ* hybridisation

Steve Bryce for help with acquiring and analysing the on-line genomic sequences

All computing, central services and library staff for help throughout

Grateful thanks for funding to:

The Cancer Research Campaign as was, and Cancer Research UK as is now, for the Gordon Hamilton-Fairley Research Fellowship for a Clinician

The University of Glasgow for a prior University Lectureship in Pathology

The Association of Clinical Pathologists and National Institutes of Health for travel grants

Thanks for making my time in MedOnc a pleasure:

O2 members past and present: Claire Anderson, Alan Bilsland, Lisa Bryce, Sharon Burns, Jayne Claire Davis, Dennis, Martin Downey, Jeff Evans, Ros Glasspool, Stacey Hoare, Katrina Jackson, Fiona McGregor, Aileen Monaghan, Jo Morris, Joe Sarvesvaran, Anja Soder, Chris Wilson and Jiangqin Zhao. For friendship, moral support, laughs, tea (in gallons, plus intermittently filched tea-bags, milk and mugs), chocolate, cakes, parties, advice on life the universe cars and bathrooms, debates on George C vs Sean B, encouragement (bullying) to attend exercise classes and five (count 'em) great years.

Brevity prevents full accountability, but extra-special thanks, for everything, to (in alphabetical order): Aileen (great shoes!); Jayne (tea and txt queen!); Katrina, magic sense of humour; Lisa (Prof Mol!); Ros, ever-patient and smiling agony aunt; Sharon, cocktail-mixer, croupier and Queen Amidala extraordinaire; and Stacey (missing S&TC already but at least we've got the glitter)

O1 members past and present: All especially Kim Appleton, Sophie Barrett, Lynne Braidwood, Scott Davidson, Steve Durant, Gill Gifford, Liz Homer, Mary McPartlin, Helen Robinson and Gordon Strathdee. For all of the above and also for education on inflatable plastic chairs, tequila slammers and of course Clatty Pat's.

Not forgetting: Debbie Stuart (Prof Cell) for friendship and discussions; Choon Seow for the same and for mutual mutterings about SAGE; and Kevin Robertson for words, wine and knickerbocker glories.

Special thanks to:

Margaret Anderson, a wonderful biology teacher, for starting me off on this track in the first place

Paul Roden for eleven great years and hopefully a lifetime's friendship: the book as promised!

Tracey Gallagher for friendship since age eleven: cheers!

My late grandad, George Morrison, who instilled in me a delight in words, writing and the classics that's dimmed only slightly over recent months

And last and definitely not least, my mum, Anne Øien: here's that thesis (of yours, and mine!) finally, with much love

Author's Declaration

I am the sole author of this thesis. All of the references have been consulted by myself in the preparation of this manuscript. Except where otherwise stated, the work presented in this thesis was performed personally.

Abbreviations

APC	adenomatous polyposis coli
APS	ammonium persulphate
BCR	breakpoint cluster region
BSA	bovine serum albumin
B+W	binding and washing (buffer)
CDK	cyclin-dependent kinase
cDNA	complementary deoxyribonucleic acid
CGAP	Cancer Genome Anatomy Project
CI	chromosomal instability
CML	chronic myeloid leukaemia
DAB	diaminobenzidine
dbEST	database of expressed sequence tags (GenBank)
DCC	deleted in colon cancer
DD	differential display
DDD	digital differential display
dH_2O	de-ionised and distilled water
DMF	dimethylformamide
DMSO	dimethylsulphoxide
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide
DTT	dithiothreitol
EBI	European Bioinformatics Institute
EDTA	ethylenediaminetetra-acetic acid
EGF	epidermal growth factor
EGFR	epidermal growth factor receptor
EST	expressed sequence tag
FBS	foetal bovine serum
FISH	fluorescent *in situ* hybridisation
FTP	file transfer protocol
GIST	gastro-intestinal stromal tumour
H&E	haematoxylin and eosin
HGMP	Human Genome Mapping Project
HGMP-RC	Human Genome Mapping Project Resource Centre, Hinxton, Cambridge, UK

HNPCC	hereditary non-polyposis colorectal cancer
Hsp90a	heat shock protein 90 alpha
IgA	immunoglobulin alpha
IHC	immunohistochemistry
IM	intestinal metaplasia
IPGT	isopropylthio-β-D-galactoside
ISH	*in situ* hybridisation
LB	Luria broth
LOH	loss of heterozygosity
LoTE	tris, ethylenediaminetetra-acetic acid (low concentration)
mRNA	messenger RNA
MCC	mutated in colon cancer
MI	microsatellite instability
M-MLV	Moloney Murine Leukaemia Virus
MUC	mucin
NSAID	non-steroidal anti-inflammatory drug
N	normal stomach
N	nodal (lymph nodal) stage in TNM staging system
NCBI	National Center for Biotechnology Information
OGJ	oesophago-gastric junction
Oligo(dT)	oligonucleotide-deoxythymidine
OMIM	On-line Mendelian Inheritance in Man^{TM}
P/C	phenol/chloroform
PAC	plasmid artificial chromosome
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PEG	polyethylene glycol
Poly(A)	polyadenylation (tail)
PSCA	prostate stem cell antigen
RACE	rapid amplification of cDNA ends
RER+	replication error positive
RNA	ribonucleic acid
RT-PCR	reverse-transcription polymerase chain reaction
RT-RDA	reverse-transcription representational difference analysis
SAGE	serial analysis of gene expression
SDS	sodium dodecyl sulphate

SEQNET	SEQuence NETwork computer for molecular biologists at HGMP-RC
SH	subtractive hybridisation
T	tumour stage in TNM staging system
T1	tumour sample 1 for SAGE
T2	tumour sample 2 for SAGE
Ts	both tumour samples for SAGE
TAE	tris, acetic acid and ethylenediaminetetra-acetic acid
TBE	tris, boric acid and ethylenediaminetetra-acetic acid
TCA	trichloroacetic acid
Tcf-4	T cell factor-4
TE	tris and ethylenediaminetetra-acetic acid (standard concentration)
TEMED	tetramethylenediamine
TFF	trefoil factor
TGFalpha	transforming growth factor alpha
TNM	Tumour, Node, Metastasis clinical staging system for cancer
Tris	2-amino-2-(hydroxymethyl)propane-1,3-diol
UV	ultraviolet
X-gal	5-bromo-4-chloro-3-indolyl-b-D-galactoside

Units

bp	base pair
°C	degree Celsius
Ci	Curie
cm	centimetre
Da	Dalton
g	gram
g	gravity
h	hour
k	kilo
l	litre
M	mega
m	metre
m	milli
M	molar
min	minute
n	nano

rpm	revolutions per minute
RT	room temperature
s	second
U	unit
μ	micro
w/v	weight for volume

Policy on references

This research described in this thesis investigates the entire spectrum of genes expressed in normal stomach and in gastric carcinoma. The number of potentially relevant published papers is therefore vast, and those included have been carefully selected.

In the materials and methods, results and discussion sections, almost all of the cited literature is primary. However, in the introduction, there is a large amount of relevant work, much of which is well-established, and some of which has changed little for decades; for this background material, most of the references are to review papers and books. Obviously, where the research described is either recent or fundamental to the thesis, however, the primary paper is cited.

1 Introduction

In this introduction, I first describe the traditional histopathological assessment of cancer, its role in patient management, and the potential contribution of new molecular pathology techniques. The normal structure and function of the stomach are then descibed. These explain its susceptibility to disease states, such as *Helicobacter pylori*-associated chronic gastritis, which are briefly discussed; conditions which are regarded as pre-malignant are highlighted. This leads naturally on to gastric carcinoma, for which the epidemiology, aetiology and pathogenesis, pathology, clinical investigation and treatment are described in detail.

The focus then moves to the molecular biology of cancer. Traditional candidate gene approaches to the identification of cancer-associated genes are explained. The tumour suppressor genes and oncogenes identified as playing a role in the development of gastric carcinoma are discussed. Examples are given of the clinical use of such cancer-associated genes. The rationale behind large-scale molecular analysis of genes and their expression is then presented; the different methods available are discussed, including the technique chosen for this study, serial analysis of gene expression (SAGE). Exciting findings from recent expression profiling studies, and their possible clinical applications, are explored.

Lastly, the aims of this research project are presented: SAGE was established locally and used to profile the genes expressed in gastric carcinoma, of distal type, and in normal stomach, then a selection of candidate genes was validated in a wider panel of gastro-intestinal tumours, and one highly expressed novel gene was investigated in detail, at mRNA and DNA levels, in human and model species.

1.1 Cancer pathology

1.1.1 Traditional histopathology of cancer

Cancers are so-called because the appearance of these invasive, and thus malignant, tumours resembles the claws of a crab: the latter is the literal meaning of the word cancer in Greek (Cotran *et al.* 1994). Thus the original description of cancer was based on appearance, or morphology, which remains the main method of assessment today.

However, while all cancers are tumours, not all tumours are cancers (reviewed in (Wyllie 1992; Cotran *et al.* 1994)). A century ago, the pathologist Rupert Willis described a tumour as, “an abnormal mass of tissue, the growth of which exceeds and is uncoordinated with that of the normal tissues and persists in the same excessive manner after cessation of the stimuli which evoked the change," and this definition holds good today. Tumours are usually classified according to two features: their biological behaviour; and the cells from which the tumour has arisen. In terms of biological behaviour, the fundamental distinction is between: benign tumours, which remain localised at their site of origin; and malignant tumours, or cancers, which spread beyond their normal tissue confines and to distant sites. In general, benign tumours rarely lead to the patient's death whereas malignant ones often do.

Tumours have traditionally been assessed by histopathology in order to predict their clinical behaviour (reviewed in (Wyllie 1992; Cotran *et al.* 1994)). Gross (naked-eye) and microscopic examination are used to confirm the diagnosis and then to predict prognosis and guide therapy. The diagnosis of cancer is made according to structural, or morphological, abnormalities at the levels of the overall tissue architecture and the individual component cells and their nuclei. The presence of tumour invasion beyond its normal limiting tissue boundaries identifies the tumour as malignant, that is, as a cancer, rather than benign. The tumour must then be classified according to the normal tissue from which the tumour originates. For example, the commonest human cancers arise from epithelial cells and are named carcinomas: this group includes gastric adenocarcinoma, which is the tumour under study here.

Prediction of the likely outcome (prognosis) for the patient is made according to the grade and stage of the cancer, which may also be used to guide therapy. The grade describes how closely the tumour resembles its tissue of origin (differentiation): low-grade tumours are well differentiated, and closely resemble their normal tissue counterparts, whereas high-grade tumours are poorly differentiated and may be difficult to classify. In contrast, tumour stage describes how far the cancer has spread, both locally at its site of origin and beyond, to draining lymph nodes and distant organs, for example the liver.

This classical anatomical and histological approach to the prediction of the behaviour of tumours has been developed and used successfully over the past 150 years or so. However, it has limitations: for any single patient, the prognosis given can only be fairly broad, since some tumours may appear similar yet behave very differently in different patients. The hope is that new assays based on the molecular changes underlying cancer may yield clinically useful information beyond that which can be provided by traditional histopathology, for, for example, better prediction of the outcome and response to therapy in individual patients.

1.1.2 Molecular pathology of cancer

The definition of a tumour provided in the previous section was based on its form and function, in other words, on classical morphology and physiology. An alternative definition could be: A tumour is formed by excessive proliferation of cells as a result of an irreversible genetic change which is passed from one tumour cell to its progeny.

The study of the molecules on which cell and tissue structure and function is based lies at the other end of the technological spectrum from classical histopathology. Analysis of the many molecular changes underlying the development of cancer, discussed in detail later in this chapter in Section 1.5, has become a viable option only relatively recently. For the past three decades or so, it has been possible to study individual genes in tumour tissues. Over the past few years, new methods have become available which enable the expression patterns of thousands of genes to be investigated simultaneously (Carulli *et al.* 1998). The application of such technologies to clinical samples is providing us with molecular profiles of normal tissues and of their corresponding tumours. These novel molecular portraits must first be correlated with existing histopathological and clinical data, in order to establish base-line information, akin to a traditional atlas of histology and histopathology. Thereafter, it is to be hoped that molecular analysis will go on to provide entirely new information, for example on tumour sub-classification, mechanisms of carcinogenesis, and the identification of novel diagnostic and therapeutic targets. These developments have been reviewed widely in the recent scientific literature (Emmert-Buck *et al.* 2000; Liotta *et al.* 2000) and are discussed in detail in Section 1.7.2.

Such large-scale molecular profiling of gene expression at the mRNA level has already been applied to a number of the common cancers, including carcinomas of the breast, colon, lung, ovary and prostate which have been identified as being of special concern by the Cancer Genome Anatomy Project of the National Cancer Institute in the US (Strausberg *et al.* 2001). Carcinoma of the stomach, however, has been less well-studied, yet it is the fourth most common cancer worldwide and carries a grave prognosis (Parkin 2001). This important tumour is therefore the subject of this study.

1.2 The normal human stomach

This study aims to profile gene expression in normal stomach and in gastric carcinoma. Many of the genes which are expressed in the stomach relate to its normal form and function, so an understanding of this is important. Furthermore, normal gastric anatomy and physiology explain much of this organ's susceptibility to disease, including cancer.

1.2.1 Anatomy and physiology of the human stomach

The stomach is located in the abdominal cavity and forms part of the gastro-intestinal tract. Food enters the body via the mouth where is undergoes only preliminary mechanical breakdown via the teeth. Early digestion of carbohydrate starch is started by the action of salivary amylase. Food is transmitted to the stomach via the oesophagus, which provides mainly a means of channelling food through the thoracic cavity, without significant further digestive activity.

The stomach acts as a reservoir for the food while it is broken down by mechanical churning and by the chemical action of gastric secretions. It is worth remembering the functions of the stomach (reviewed in (Owen 1986; Tortora *et al.* 2000)). It acts as a reservoir for food and mechanically churns and mixes it with gastric juice containing: hydrochloric acid which is a sterilising agent, denatures proteins and activates digestive enzymes; the protease pepsin; intrinsic factor; and gastric lipase. The stomach is not a major absorptive site, but water, ions, short chain fatty acids and alcohol are absorbed. The stomach must also protect its mucosal lining from proteolytic and acid attack, through the production of mucus and associated proteins. The muscular and secretory activities of the stomach must be co-ordinated with the rest of the gut through neuro-hormonal mechanisms. From the stomach, food passes to the duodenum where the secretions of the small intestine and pancreas, and bile, neutralise the acidity of the gastric output and add further enzymes and detergents.

In the stomach, as elsewhere in the gastro-intestinal tract, the wall is made up of four layers: mucosa, submucosa, muscularis and adventitia (reviewed in (Lewin *et al.* 1996; Tortora *et al.* 2000)). The mucosa is innermost: it lines and secretes into the lumen. The submucosa conducts blood vessels, lymphatic channels and nerves to the mucosa and is mainly composed of fat which provides support and acts as a shock-absorber. The muscle coat (named the muscularis propria) comprises two layers in most areas of the gut: an inner circular and outer longitudinal coat. The stomach contains an additional third muscle layer, which is orientated obliquely. The muscle propels the food along the gastro-intestinal tract by peristalsis; in the stomach, with the help of the third muscle layer, the muscle provides the back-and-forth churning action which mechanically breaks down the food. In addition, specialised adaptations of the muscle coat, where it is thickened and specially orientated, called sphincters, provide valves throughout the gastro-intestinal tract which control the passage of food. The distal valve in the stomach separates it from the duodenum and is called the pylorus. At its proximal end, between the stomach and the oesophagus, a valve exists but is less well-defined and comprises a slight thickening of the muscular coat combined with surrounding slips (loops) of muscle from the adjacent diaphragm. The outermost layer of the gut wall, the adventitia, transmits blood vessels, lymphatics and nerves to the inner layers and comprises mainly thin, loose connective tissue. For most of the stomach (as opposed to, say, the oesophagus), the adventitia actually forms part of the peritoneal lining of the abdominal cavity and hence is called serosa.

Broadly speaking, the submucosa, muscle coat and adventitia are similar throughout the gut. What varies between the different organs in the gastro-intestinal tract is the mucosa, which comprises the epithelium lining the lumen, the supporting lamina propria and the underlying muscularis mucosae. The muscularis mucosae is a very thin muscle layer which helps to propel mucosal secretions into the lumen; it is not the same as the previously described muscularis propria. The lamina propria transmits blood vessels and lymphatic channels from the submucosa to the epithelium, and is the main location of the inflammation in gastritis (see Section 1.3.1). The mucosa and its epithelial lining in the different areas of the stomach are described in detail in the next section.

1.2.2 Histology of the human stomach

The stomach is divided anatomically and by its mucosa into three distinct areas (Figure 1.1), all lined by simple glandular epithelium (reviewed in (Lewin *et al.* 1996; Tortora *et al.* 2000)). The cardia is proximal and comprises the first 1-2 cm of the stomach around the entrance to the oesophagus. Its mucosa is predominantly mucus-secreting, and is thinner and less complex than those elsewhere in the stomach.

The bulk of the stomach is lined by mucosa of gastric body type which comprises long, tubular glands lined by a variety of epithelial cells (Figure 1.2). Within each gland, the upper quarter is lined by mucous (mucus-secreting) cells of surface / (glandular) neck type. The next cells form the stem cell (proliferative) compartment. The next quarter of the gland crypt contains parietal cells. These are the gastric acid-secreting cells, which secrete hydrogen and chloride ions into the gastric lumen. Because of the high ionic gradients which they must achieve across their cell membranes, parietal cells have a very high energy requirement and therefore contain large numbers of mitochondria, which yield a dense pink granularity to their microscopic appearance on routine histological staining with haematoxylin and eosin (H&E). In the lowermost quarter of the gland are the gastric chief, or peptic, cells which secrete the main gastric enzyme, pepsin, in its precursor form, pepsinogen. Since their main role is to produce a protein for secretion in abundance, these cells contain large quantities of rough endoplasmic reticulum and hence stain dark purple with H&E staining. The basal part of the glands also contains endocrine cells, which here in the gastric body are termed enterochromaffin-like cells.

The distal part of the stomach is called the antrum. Within the antrum itself, its distal portion is termed the pylorus and acts as a sphincter. Gastric antral mucosa is simpler than that in the body, being composed mainly of mucus-secreting glands (Figure 1.2). In addition, there are endocrine cells: the most numerous here in the antrum of the stomach secrete gastrin and hence are termed G-cells.

1.2.3 Neuro-hormonal control of the human stomach

Gastric digestion occurs in three overlapping phases: cephalic, gastric and intestinal (Tortora *et al.* 2000). The cephalic phase consists of reflexes initiated by the sight, smell, taste or thought of food, which activate the vagus (tenth cranial) nerve. Via submucosal parasympathetic fibres, secretions from parietal, chief, mucous and G-cells are stimulated, as is gastric motility.

When food enters the acid environment of the stomach, the luminal pH rises from its normal of around pH 1 and the stomach physically expands. Chemoreceptors and stretch receptors further stimulate gastric secretions and motility in what is called the gastric phase and also cause the G-cells in the gastric antrum to secrete the hormone gastrin into the bloodstream. Gastrin and vagal impulses further promote secretion of gastric acid and pepsin(ogen), strengthen contraction of the proximal sphincter at the oesophago-gastric junction (OGJ) and relax the distal, pyloric sphincter.

As the food is mechanically and biochemically broken down to so-called chyme and gradually passes through the pylorus into the duodenum and beyond, digestion enters the intestinal phase. Receptors in the small intestine provoke the local release of two more hormones, secretin and cholecystokinin, which inhibit gastric secretions and emptying, preventing overloading of the duodenum with chyme. In addition, as the stomach empties and its low pH is restored, further negative feedback loops operate to inhibit vagal signals and gastrin release.

1.3 Gastric disease including gastric pre-neoplasia

Disease may be defined as any deviation from normal structure or function. Within the stomach, most disease relates to the harsh luminal environment containing acid, proteases and foreign material, combined with occasionally inadequate mucosal protection. Because the subject of this project is gastric carcinoma of distal type (see Section 1.4) and corresponding normal mucosa of the distal stomach (antrum, and, to a lesser extent, body), the main focus is on gastritis; proximal disease, of the gastro-oesophageal junction, is discussed only briefly.

1.3.1 Distal stomach: gastritis and Helicobacter pylori infection

Within the main portions of the stomach, the gastric body and antrum, the most common disease process is gastritis. Gastritis literally means inflammation of the stomach and inflammation may be simply defined as the body's response to injury (Cotran *et al.* 1994).

Classically, the list of injurious agents causing gastritis has included unusually harsh foodstuffs such as alcohol and excessively spiced, salted or smoked foods, and drugs especially the non-steroidal anti-inflammatory drugs (NSAIDs) such as aspirin (Cotran *et al.* 1994; Calam *et al.* 2001). Whilst these have previously been considered to be toxic mainly through direct, local mechanisms, it is now thought that, for aspirin at least, its systemic effects on, for example, mucosal cyclooxygenase enzymes may also play a role. Reflux from the small intestine of alkaline secretions, including bile, is also directly toxic and can cause gastritis: this is common after gastric surgery resulting in anatomically abnormal links with the small intestine (i.e. a gastro-enterostomy) which effectively removes the one-way valve effect of the pylorus.

Other traditional causes of gastritis involve mechanisms whereby the gastric mucosa is less able to protect itself (Cotran *et al.* 1994). These include stress, especially during illness or after surgery, steroid therapy and smoking, which may alter the gastric vasculature. In addition, having Blood Group O confers different antigenic properties on the mucins (Lewis antigens) expressed by gastric epithelium and thus alters their susceptibility to attachment and invasion by infectious agents, and hence predisposes to gastritis.

The list of causes of gastritis given above is similar to that given in a book entitled "Gastric and Duodenal Ulcer" published in 1929 (Hurst *et al.* 1929), but the most exciting development in our understanding of gastro-intestinal disease for decades has been the identification of *Helicobacter pylori* and its recognition as by far the major cause of gastritis worldwide and thence, through gastric atrophy, of distal gastric carcinoma.

Helicobacter pylori is a gram-negative bacterium, with a spiral (helical) shape and an ecological niche limited to the stomach, hence the name. *H. pylori* was first described in 1984 (Marshall *et al.* 1984). Recent reviews of *H. pylori* infection cover its epidemiology (Logan *et al.* 2001), pathophysiology (Calam *et al.* 2001) and clinical management (Harris *et al.* 2001). It is the most common chronic bacterial infection in the world, infecting around 50% of the population. Its prevalence is linked to social deprivation, being highest in the third world and South-East Asia. In the Western world, it is most common in people of lower socio-economic status. Transmission is mainly in childhood and is thought to be faecal-oral. Most people infected with *H. pylori* are asymptomatic, but a substantial minority complain of dyspeptic symptoms, including abdominal pain, bloating and nausea. The long-term course of *H. pylori* infection over years also varies. In some people, the infection is eradicated early, either spontaneously or through the use of antibiotics, often for an unrelated indication. In most, the infection persists for many years.

Long-standing gastritis due to *H. pylori* infection varies between individuals in the area of the stomach principally affected, the relative degree of inflammation, the functional results and the associated symptoms (Blaser 1998; Calam *et al.* 2001). Over 70% of individuals infected with *H. pylori* are asymptomatic. Most patients have a gastritis which is histologically most marked in the antral mucosa and which is predominantly chronic, that is, containing chronic inflammatory cells, especially lymphocytes and plasma cells. Most individuals with an antral-predominant gastritis have normal gastric acid levels and are asymptomatic. Others, however, have raised gastric acid; these patients are more prone to be symptomatic and, in particular, to the development of duodenal ulcers (Calam *et al.* 2001). These are caused by excess acid entering the duodenum and damaging its mucosa, which is not, unlike the gastric mucosa, acid-adapted, hence the alternative name of peptic ulcer.

In contrast, a minority of infected individuals have a body-predominant gastritis (El-Omar *et al.* 1997). Since the gastric body is the site of acid secretion, their gastric acid levels are usually low (and thus the gastric luminal pH is high, approaching neutral), especially if the gastritis is accompanied by atrophy, as discussed in the next paragraph (El-Omar *et al.* 1997). These individuals are usually asymptomatic but appear to be more likely to develop gastric ulcers, which are also related to other causative factors such as NSAIDs. More importantly, these are the *H. pylori*-infected patients who are especially prone to developing gastric carcinoma (El-Omar *et al.* 2000).

Long-standing gastritis, whether in the body or antrum, often leads to further changes in the gastric mucosa: intestinal metaplasia (IM) and glandular atrophy (Figure 1.2) (Lewin *et al.* 1996). Metaplasia can be defined as the change of one differentiated cell type into another differentiated cell type. Native gastric mucosa has a characteristic appearance, as described in Section 1.2.2: the surface and foveolar (pit) epithelium is similar throughout and is mucus-secreting; the glands in the antrum are also produce mucus, whereas those in the body contain acid-secreting parietal cells and pepsinogen-producing chief cells. Intestinal metaplasia results in a change to part or all of the gastric glandular thickness: the surface epithelium then comprises intestinal absorptive cells with a surface microvillous brush border, interspersed with goblet cells, and the glands contain a proliferative compartment with basal anti-bacterial Paneth cells and endocrine cells (Figure

1.2). The metaplastic intestinal epithelium provides a rather hostile environment, which does not secrete acid nor support the growth of *H. pylori.*

Gastric glandular atrophy often accompanies intestinal metaplasia and can be defined as a reduction in the number and size of the gastric glands (Figure 1.2), due to chronic inflammation and scarring (Lewin *et al.* 1996). Intestinal metaplasia and mild degrees of atrophy in the gastric antrum are common with chronic *H. pylori* infection but alone are usually of little functional significance. Intestinal metaplasia and any degree of atrophy in the gastric body are usually a consequence of body-predominant gastritis, and like that condition, are associated with low acid secretion, which is obviously of considerable functional importance (El-Omar *et al.* 1997). Gastric atrophy, or atrophic gastritis, in the body mucosa is usually the end-result of *H. pylori* infection but another cause is auto-immune gastritis, also known as pernicious anaemia. Both intestinal metaplasia and atrophy are regarded as pre-malignant conditions for gastric carcinoma (of distal location and intestinal histological type), to which they may progress, sometimes via the intermediary step of epithelial dysplasia (Correa 1988; Correa 1992; Uemura *et al.* 2001) (see Section 1.4.6).

1.3.2 Proximal stomach: acid reflux and Barrett's oesophagus

The discussion above has concentrated on gastritis of the body and antrum and the sometimes associated duodenal ulceration, which until recent years were by far the most prevalent type of upper gastro-intestinal non-neoplastic diseases. However, over the past two decades, disease of the gastro-oesophageal junction has become as common and important. Problems at this site relate to reflux of gastric, and possibly duodenal, contents into the oesophagus, which causes oesophagitis (inflammation of the distal oesophagus), ulceration (loss of the surface epithelium) and repair in the form of yet another metaplastic epithelium named Barrett's oesophagus (Lewin *et al.* 1996).

In Barrett's metaplasia, the normal squamous epithelium lining the oesophagus is replaced by glandular epithelium of both intestinal and gastric (cardiac and sometimes body) types. The injury is thought to be caused mainly by the acid within the refluxing fluid, although bile and other duodenal contents may also contribute, and so its treatment involves agents which lower gastric acid secretion, such as proton pump inhibitors e.g. omeprazole. As is the case in the distal stomach and elsewhere in the body, such as in the bronchial tree and female cervix, this metaplastic epithelium is at greater risk than the native mucosa of undergoing dysplastic and eventually malignant change, developing into a Barrett's-associated adenocarcinoma (Lewin *et al.* 1996).

1.4 Gastric carcinoma

Now that the normal structure and function of the stomach, and its associated non-neoplastic diseases have been described, we can move to a discussion of gastric carcinoma, about which recent reviews and texts include (Campaign 1995; Fuchs *et al.* 1995; Lewin *et al.* 1996; Blok *et al.* 1997; McAleer 2001).

1.4.1 The most common cancer of the stomach is adenocarcinoma

Over 90% of cancers in the stomach are carcinomas (more specifically, adenocarcinomas), that is, they are malignant tumours derived from surface epithelium (carcinomas), and the prefix "adeno-" is applied because the epithelium is of glandular type. These are the topic of this research but other gastric cancers exist: most are either lymphomas or stromal (connective tissue) tumours (Lewin *et al.* 1996).

Gastric lymphomas tend to be of MALT (mucosal-associated lymphoid tissue) type. Like distal adenocarcinomas, they are associated with *H. pylori* infection; the lymphoma develops from a monoclonal B-lymphocyte population arising within the chronic inflammatory infiltrate of gastritis (Lewin *et al.* 1996). Many gastric MALT lymphomas are low-grade and can be cured simply by eradication of *H. pylori* with triple therapy using antibiotics and acid suppression (Blaser 1998).

Our understanding of gastro-intestinal stromal tumours (GISTs) has advanced considerably over recent years. Previously stromal tumours were thought to be derived mainly from either smooth muscle or peripheral nerves, that is, they were either leiomyosarcomas or schwannomas (Lewin *et al.* 1996). It is now known that most GISTs arise from the interstitial cells of Cajal, which are cells of the myenteric plexus, involved in neuro-hormonal control of gastric function. GISTs specifically over-express the oncogene c-kit (also known as CD117) because of an activating mutation. c-kit can now be successfully targeted with the small molecule inhibitor STI571: this is one of the recent success stories of molecular genetic therapy (Mauro *et al.* 2002) and is further described in Section 1.5.6.

1.4.2 Gastric carcinoma may be of distal or proximal location

Until recent years, throughout the world, almost all oesophageal carcinomas were of squamous type, and most gastric adenocarcinomas were located in the gastric antrum or body (regarded as distal) (Lewin *et al.* 1996). In contrast, proximal adenocarcinomas of the oesophago-gastric junction (OGJ) were a rarity. However, proximal tumours have greatly increased in incidence over the past two decades, such that they are now as common as distal gastric carcinomas, and more common than oesophageal squamous carcinomas (Fuchs *et al.* 1995; Blok *et al.* 1997). OGJ adenocarcinomas are usually taken to include both proximal gastric tumours located in the cardia and lower oesophageal adenocarcinomas, often arising in a background of Barrett's oesophagus; even here, there is some evidence that the cardiac and oesophageal tumours should be considered separately, but their accurate sub-division is often difficult at both operation and pathological examination.

Gastric adenocarcinomas therefore must be considered in (at least) two groups according to their anatomical location: distal and proximal. The following discussion of gastric adenocarcinoma covers epidemiology, pathology, clinical presentation, outcome and treatment, and the molecular biology will be described later in Section 1.5.4. In most cases, the discussion is similar for the two tumour sites but where differences exist, the focus will be on the distal tumours which are the subject of this research.

1.4.3 Epidemiology

Gastric carcinoma as a whole is a common tumour with a high mortality (Fuchs *et al.* 1995; Blok *et al.* 1997). In 2000, stomach cancer was the fourth most frequent cancer worldwide, with 876 000 new cases (8.7% of the total) and 647 000 deaths (10.4% of cancer deaths) (Parkin 2001). It was surpassed by cancers of the lung, breast and colon. Almost two-thirds of these cases occurred in less developed countries. Age-standardised incidence rates are highest in Japan (69.2 per 100 000 in men, 28.6 per 100 000 in women) (Parkin 2001). High rates are also present elsewhere in eastern Asia and in Central and South America and eastern Europe.

In the UK, there are still around 11,000 new cases annually and stomach cancer was the fifth most common cause of cancer death in 1999 (Campaign 2001), with similar rankings in the rest of Europe and in the US. The five-year survival in the UK is around 10%. This figure is comparable to carcinoma of the oesophagus, much better than that for pancreatic carcinoma (2%), and much worse than those for colorectum (40%) and breast (74%) (Campaign 2001). Throughout the world, gastric carcinoma as a whole is more common in older age groups, increasing markedly after the sixth decade, as is the case with most adenocarcinomas, in keeping with a usually sporadic origin due to accumulated genetic events (see Section 1.5). The disease is more common in men than women with a ratio of around 3:2. In the Western world, the tumour is more common in lower socioeconomic groups (Fuchs *et al.* 1995; Blok *et al.* 1997). The epidemiology of gastric carcinoma thus parallels exactly the pattern of *H. pylori* infection.

As stated in the previous section, the incidence of gastric carcinoma has declined worldwide over recent decades: in 1980, it was the second most common cause of cancer death, and it was the leading cause until 50 years ago (Fuchs *et al.* 1995; Blok *et al.* 1997). This is due to fewer tumours of the distal stomach, which are the most common in areas of high incidence. The decrease has occurred in the absence of any major changes in diagnosis or therapy, and likely causes are discussed in Section 1.4.6. In contrast, proximal gastro-oesophageal adenocarcinoma is becoming more common in developed countries, including the UK and USA. Its rise has been precipitous and is of a similar gradient to that of malignant melanoma in the 1980's (Fuchs *et al.* 1995; Blok *et al.* 1997). The reasons are unclear but possible causes are again debated in Section 1.4.6.

1.4.4 Gross pathology

Since gastric carcinomas are derived from epithelium, they arise in the mucosa from where they invade longitudinally, circumferentially and deeply, into the submucosa and underlying muscle wall (reviewed in (Cotran *et al.* 1994; Lewin *et al.* 1996; McAleer 2001)). From there, the tumour may directly invade adjacent organs such as the pancreas posteriorly and transverse colon anteriorly. By gross (naked-eye) examination, most gastric carcinomas are either ulcerating lesions, often with raised rolled edges, or nodular polypoid tumours (the latter is illustrated in Figure 1.3). Some carcinomas, however, are mainly submucosal where they infiltrate diffusely causing a markedly thickened stomach wall with a reduced capacity, described as linitis plastica, literally meaning leather-bottle stomach. Of the true gastric tumours, half arise in the pyloric region with most of the remainder along the lesser curvature (gastric anatomy is described in Section 1.2.1).

Tumour spread within lymphatics is common because the stomach and oesophagus, unlike the colon, for example, contain abundant mucosal and submucosal lymphatic channels (Lewin *et al.* 1996). This explains the frequency of early lymphatic spread in gastro-oesophageal carcinoma and also explains why it is unfortunately common to find microscopic deposits of tumour at the resection margins followng surgery. Lymphatic spread is first to local draining lymph nodes in the lesser and greater omentum (see Figure 1.3) and thence to regional lymph nodes in the coeliac plexus.

Spread in a transcoelomic (across body cavities) manner gives rise to peritoneal deposits which may be very large: clinically this is described as “omental cake”. The same mechanism may also result in the well-recognised clinical phenomenon of the Krukenberg tumour, comprising bilateral metastatic deposits in the ovaries in young women. Haematogenous spread may occur through drainage via the portal circulation from the gut to the liver or via the systemic circulation to the usual sites of lung, brain and bone.

The tumour stage describes the extent of local and distal tumour spread, and the current TNM ((local) Tumour, (lymph) Node, Metastasis) staging system for gastric carcinoma is described in Table 1.1 (Klein Kranenbarg *et al.* 2001). The concept of "early" cancer is peculiar to the stomach and constitutes tumour confined to the mucosa or submucosa, regardless of the presence or absence of lymph node spread (Lewin *et al.* 1996). This has a much better prognosis than tumour which has spread into the muscle coat or beyond. This implies that for gastric adenocarcinoma at least, transcoelomic or haematogenous metastases carry a worse prognosis than spread to local lymph nodes.

1.4.5 Histopathology

The subdivision of gastric adenocarcinomas is not only by anatomical location (Fuchs *et al.* 1995; Lewin *et al.* 1996; Blok *et al.* 1997). In addition, there are different histological tumour subtypes, with the most common classification separating tumours into intestinal and diffuse cancers (Laurén 1965). Not surprisingly, some are of mixed type (Laurén 1965) and those arising at the oesophago-gastric junction may show some squamoid features. As previously described, the tumours then undergo histological grading in terms of how closely the tumour resembles its tissue of origin (differentiation).

"Intestinal" tumours comprise cohesive cells forming glandular structures, as illustrated in Figure 1.3, and arise mainly distally in an elderly population. Since the glands of "intestinal" carcinomas recapitulate the normal gastric mucosal architecture, these tumours tend to be well or moderately differentiated. As a gross generalisation, the epidemiology of intestinal gastric adenocarcinoma broadly parallels that of distal tumours: in high incidence areas, intestinal gastric carcinomas predominate and their reduction is considered to be responsible for the recent decline in gastric cancer rate.

"Diffuse" cancers contain individual, infiltrating malignant cells, as seen in Figure 1.3; they may develop at any site in the stomach but especially in the cardia, and carry a worse prognosis. The individually infiltrating cells of "diffuse" cancers bear almost no resemblance to normal stomach and therefore they are usually regarded as poorly differentiated *ab initio*. Diffuse cancers have a greater propensity for serosal spread. Their incidence, in contrast to that of intestinal cancers, is similar and apparently static worldwide (Campaign 1995; Fuchs *et al.* 1995).

It is worth noting that some authors have suggested that the pathological evaluation of gastric carcinoma, and possibly the disease itself, may vary between Japan and Western countries (Schlemper *et al.* 1997). In Japan, cancers of distal, intestinal type predominate. The criteria used to define cancer differ: in the West, pathologists need to see invasion, but in Japan, abnormal cellular morphology is sufficient. This diagnostic dispute matters only for early stage disease: what would in the West be called severe dysplasia would be called intramucosal carcinoma in Japan. However, for cancers involving submucosa and beyond, the diagnostic criteria are similar worldwide and therefore comparable.

Overall, then, gastric carcinoma, whether categorised by anatomical site or by histological type, can be regarded as two (or more) diseases which differ epidemiologically, clinically and morphologically, as well as in their pathogenesis and molecular biology, as discussed next and later in Section 1.5.4.

1.4.6 Aetiology and pathogenesis

The marked variation in worldwide incidence, together with migrant studies, suggest the importance of environmental factors in the development of distal gastric carcinoma of intestinal type (reviewed in (Campaign 1995; Blok *et al.* 1997; McAleer 2001)). Recognised dietary risk factors include a high intake of smoked and salted foods and nitrate, and low consumption of fresh fruits and vitamin C. Distal gastric tumours are also associated with low socio-economic status. The same is true of low acid states, including pernicious anaemia (autoimmune gastritis) and after gastric surgery such as partial gastrectomy or gastroenterostomy, which lead to chronic reflux of bile salts into the stomach, and hence gastritis, as well as to lower levels of secretion of antral-derived gastrin which normally stimulates acid secretion. Blood group A is also linked to gastric cancer; like blood group O's association with peptic ulcer, the effect probably relates to the resulting variant mucin antigens expressed by the gastric epithelium and hence altered susceptibility to bacterial attachment and infection.

The likely pathway for distal gastric carcinogenesis was originally proposed by Pelayo Correa and leads from chronic gastritis, via intestinal metaplasia and gastric atrophy, eventually progressing, possibly via dysplasia, to cancer (Correa 1988; Correa 1992). Each of these stages is histologically identifiable, but the changes, although distinct, are usually multifocal and usually multiple stages are present simultaneously.

Within the past decade, *Helicobacter pylori* infection, described previously in Section 1.3.1, has been causally implicated in at least the early stages (Group 1993; Solcia *et al.* 1996; Uemura *et al.* 2001). While almost all patients who develop distal gastric carcinoma of intestinal type have had *H. pylori* infection, the tumour occurs in less than 1% of *H. pylori*–infected individuals. The risk of cancer is associated mainly with gastritis which is body-predominant and accompanied by atrophy (El-Omar *et al.* 2000; Uemura *et al.* 2001). Despite the eventual reduction in the number of glands, gastritis *per se* is associated with increased proliferation of the mucosal epithelial cells and hence greater susceptibility to mutational events (Blaser 1998; Calam *et al.* 2001), as discussed later in Section 1.5.1. The inflammatory infiltrate is also rich in reactive oxygen species which are also potentially DNA-damaging. The atrophy results in low gastric acid levels, which permit the stomach to be colonised by anaerobic non-*H. pylori* bacteria (Correa 1992). While *H. pylori* is usually necessary for distal gastric carcinogenesis, it is not sufficient. Rates of colonization by *H.pylori* are similar in Japan and the UK, yet the incidence of gastric carcinoma is lower here than in the Far East: other factors, especially dietary, are thus implicated.

In low acid states, the abnormal colonising bacteria are able to convert excessive dietary nitrates, nitrites and secondary amines to carcinogenic nitrosamines (Correa 1988; Correa 1992). Such N-nitroso compounds have been shown to cause gastric tumours in animals. Their chemical conversion is promoted when the diet is low in anti-oxidants such as Vitamin C. Nitrates and nitrites were previously used to preserve meat, fish and vegetables. During this century, however, the nitrate and nitrite content of food in the Western world has declined by 75%, associated with increased use of refrigeration and other improved methods for preserving food. Other dietary and luminal factors include excessive salt intake and biliary reflux, which cause gastric mucosal damage and atrophy, and smoked foods which contain further carcinogens, such as benzpyrenes (Correa 1988; Correa 1992).

Intestinal metaplasia (IM) has been described above and previously in simple terms as a single entity (Section 1.3.1), but in fact there is more than one type, associated with different cancer risks (Lewin *et al.* 1996). Type I IM resembles normal small intestinal mucosa, with epithelium composed entirely of small intestinal absorptive cells with intervening goblet cells (Figure 1.2). Type II IM is different and comprises intestinal-type goblet cells interspersed with gastric-type mucous cells (Lewin *et al.* 1996). Type II IM in turn has been subdivided accoding to its mucin content, with Type IIa containing small intestinal and gastric type mucins whereas Type IIb (also, confusingly, called Type III) IM contains colonic type sulphomucins (Lewin *et al.* 1996); these differences can be demonstrated by special staining for mucins. Type IIb IM in particular is associated with an increased risk of developing distal gastric adenocarcinoma. Unfortunately although the association is strong, Type IIb IM is also relatively common, so its positive predictive value, at least in Western populations, is low and it therefore cannot be used effectively to screen for gastric cancer or to determine endoscopic follow-up (Lewin *et al.* 1996).

The next step on from IM in Correa's pathway is dysplasia, in which the epithelial cells start to develop cytological features of malignancy, such as increased nuclear size and altered nuclear shape, but remain confined within the epithelial basement membrane. Dysplasia is commonly found immediately adjacent to invasive tumours in resection specimens (Lewin *et al.* 1996). However, dysplasia in the distal stomach in the absence of carcinoma, as a true precursor lesion, is very rare in Western populations, so cannot be used as a screening tool. In areas of high gastric cancer incidence, however, such as Japan, dysplasia as a precursor is much more common and often occurs in the form of adenomas. Adenomas are polypoid growths (grossly resembling small cauliflowers) of the gastric mucosa, with dysplastic epithelium. By definition adenomas are benign since the epithelium is not invasive, but they do have malignant potential, much like colonic adenomas (Fearon *et al.* 1990).

Overall, then, the distal, intestinal type of gastric cancer is related largely to environmental factors prevalent early in life. Exposure to *H. pylori* infection and a diet deficient in fruit and vegetables and rich in highly salted or poorly preserved foods may lead to gastritis and atrophy. Further mucosal injury by intraluminal bacteria, bacterial activation of procarcinogens, or consumption of other carcinogens may lead to the development of intestinal metaplasia, dysplasia and ultimately carcinoma. Consequently, the worldwide decline in this type of gastric cancer may be the result of the diminishing prevalence of these environmental factors brought about through improved socio-economic conditions with better food storage and reduced transmission of *H. pylori.*

For gastric carcinoma of diffuse type, the risk factors are largely unrecognised, except for E-cadherin mutations (see Section 1.5.5 later); its incidence is similar worldwide and has remained largely static over the past decades (Fuchs *et al.* 1995; Blok *et al.* 1997).

For proximal adenocarcinomas, those arising in the lower oesophagus are clearly associated with Barrett's oesophagus, where the normal lower oesophageal squamous epithelium is replaced by metaplastic columnar mucosa, and in which dysplasia and carcinoma may supervene (Lewin *et al.* 1996). The rising incidence of proximal gastro-oesophageal tumours appears to be strongly correlated with a concomitant increase in Barrett's metaplasia (Fuchs *et al.* 1995; Blok *et al.* 1997). The cause of the latter is unclear although the rise in Barrett's oesophagus and associated adenocarcinoma exactly opposes the fall in distal gastric cancer and *H. pylori* infection (Blaser 1998), with the increase most marked in white males of high socio-economic class. On average, *H. pylori* infection in a population tends overall to lower gastric acid levels, although obviously certain individuals have acid excess. Its eradication has led to increases in average gastric acid levels; within fluid refluxing into the oesophagus, this could increase mucosal damage. Bile within the refluxate and excessive dietary fat, and possibly other carcinogens, may also play a role. This recent phenomenon is as yet poorly understood but much studied.

1.4.7 Clinical presentation, diagnosis, prognosis and treatment

Most patients with gastric carcinoma in the Western world develop symptoms and present to their doctor when local disease is advanced, often with metastases (reviewed in (Fuchs *et al.* 1995; Blok *et al.* 1997; Roukos 2000; McAleer 2001)). This insidious presentation occurs because the stomach is hugely distensible, through its normal role as a food reservoir, and can maintain some degree of function even when it contains a large tumour. The minimal gastro-intestinal symptoms may include nausea, vomiting, anorexia and weight loss, epigastric discomfort or bloating. Chronic blood loss may cause anaemia. Distal tumours may cause gastric outlet obstruction. In contrast, proximal tumours at the OGJ may cause oesophageal obstruction; oddly enough, despite the relatively narrow oesophageal lumen, these tumours still tend to present at a late stage. The clinical signs may include a palpable epigastric mass, and, rarely but classically, lymphatic spread may cause enlargement of the lymph nodes in the left supraclavicular fossa (Troisier's sign).

Given such upper gastro-intestinal symptoms, the clinical investigation of choice is endoscopy. This enables visualisation of the oesophagus, stomach and duodenum, with biopsy of any suspicious lesions allowing an initial diagnosis of gastric carcinoma. Thereafter, the definitive treatment, and currently the only hope for cure, is surgery.

Before radical resection is undertaken, and in the absence of clinically overt metastatic disease, further investigation is required to exclude the presence of occult metastases (M staging), in the liver, peritoneal cavity or para-aortic lymph nodes. This usually includes CT (computerised tomogram) scanning of the abdomen and chest, plus laparoscopy with peritoneal cytology (Tschmelitsch *et al.* 2000). The extent of disease locally in the stomach and in regional lymph nodes (T and N staging) is assessed simultaneously, and in some centres endoscopic ultrasound is also used.

For surgery with curative intent, partial or total gastrectomy is needed, depending on the site and size of the tumour. In Japan, where gastric carcinoma is much more common and where it is often detected at an earlier stage, the surgery is more rigorous and certainly involves more extensive lymph node dissection (Roukos 2000). Stage-for-stage, the resulting survival rates are significantly better than those reported in patients from Europe or the USA, as shown in Table 1.2 (Fuchs *et al.* 1995). For palliation (that is, where cure is not possible), especially where luminal obstruction has occurred, the less radical surgical procedures of partial gastrectomy or bypass gastroenterostomy may be useful in alleviating local symptoms. In patients unfit for surgery, endoscopic procedures such as laser ablation or oesophageal stenting may relieve oesophageal or gastric outlet obstruction.

Currently the only curative treatment for gastric carcinoma is surgery, since effective adjuvant therapies are not yet well-established (Roukos 2000). However, gastric resection intending to cure succeeds in less than 40%. Unfortunately, 40-65% of patients suffer loco-regional recurrence: in the gastric or tumour bed; at the anastomosis; or in regional lymph nodes. The frequency of such relapse makes regional radiation an attractive possibility for adjuvant therapy, but gastric carcinoma is relatively radio-resistant, requiring doses that exceed the tolerance of surrounding structures such as the bowel and spinal cord, if adequate control of the primary tumour is the aim. Thus for patients with locally recurrent or metastatic diseae, moderate doses of external-beam irradiation are currently used only to palliate local symptoms, such as dysphagia, haemorrhage or pain, and not to improve survival (Fuchs *et al.* 1995), but clinical trials continue.

For patients with surgically resectable gastric carcinoma, it was thought until recently that post-operative (adjuvant) chemotherapy or chemo-radiotherapy offered no survival advantage. However, studies over the past decade have suggested the opposite (Shimada *et al.* 1999; Macdonald *et al.* 2001; Valle 2001). The use of adjuvant therapy has now been advocated for patients who are either lymph-node positive or who have locally advanced disease (tumour stage T2 or over, i.e. not early gastric cancer), but the optimum regimen, possible neo-adjuvant use, and role of newer agents are still to be determined (Valle 2001), within the setting of clinical trials.

In advanced gastric cancer, the most effective single agent chemotherapy is 5-fluorouracil, which has been used in various combinations with doxorubicin, cisplatin, epirubicin and methotrexate: these show response rates of around 50% and prolong survival by comparison with best supportive care. In the UK, the standard combination is ECF, containing epirubicin, cisplatin and 5-fluorouracil, but different regimens are used elsewhere (Shimada *et al.* 1999).

The overall 5-year survival for gastric carcinoma in Scotland is only around 12%, which compares poorly with 44% for colonic carcinoma (Harris *et al.* 1998). The UK-wide outlook is similarly poor at around 10% (Campaign 2001), because most patients (at least 80%) with gastric carcinoma in the Western world are diagnosed late, when curative surgical resection is no longer a prospect. Furthermore, the increasingly common proximal gastric carcinomas have an even poorer prognosis than distal tumours (Fuchs *et al.* 1995). Much better 5-year survival figures of over 80% can be achieved, however, when the tumour is identified and treated at an earlier clinical stage (Roukos 2000). As the Cancer Research Campaign's Factsheet on UK Stomach Cancer commented, "The challenge is to increase the number of patients in this (early diagnosis) category". But how is this to be achieved?

In Japan, where the disease incidence is approximately five-fold higher, earlier diagnosis and corresponding improved survival have been achieved by widespread endoscopic screening and aggressive surgical management (Roukos 2000). As previously discussed, some authors have suggested that the pathological evaluation of gastric carcinoma, and possibly the disease itself, may differ between Japan and Western countries (Schlemper *et al.* 1997). Regardless, the lower UK rates of gastric cancer make similar mass population screening unjustified on cost-benefit grounds (Campaign 1995). More worthwhile strategies against gastric cancer in this country would include: earlier diagnosis by identification and monitoring of high risk groups; improved surgery and adjuvant therapy; and, finally and ideally, prevention by intervention where risk factors, such as *Helicobacter pylori* infection (see Section 1.4.6), are known (Campaign 1995). Clinical trials of more radical surgery, chemoradiotherapy, and population-based *H. pylori* eradication are already underway in the UK, but effiective markers for early diagnosis are still awaited.

1.4.8 Historical note on our knowledge of the stomach and its associated cancer

It is salutary to note that the contents of most of the above text on the normal anatomy and functions of the stomach, and on the epidemiology, pathology, aetiology, pathogenesis and treatment of gastric cancer would have been very similar twenty or even fifty years ago. Perhaps the only major advance in these themes has been the identification of *H. pylori* as a major cause of gastro-duodenal disease and of gastric cancer in particular. Its eradication would almost certainly be a major contribution to the prevention of distal gastric carcinoma worldwide: such large studies are already underway. (Whether this might then lead to an increase in the incidence of Barrett's oesophagus and its associated proximal gastro-oesophageal adenocarcinoma is a different question (Blaser 1998).) This is very different from the situation for malignant lymphoma, or for breast or colonic carcinoma, in which better appreciation of the mechanisms and molecules involved in their development has led to earlier diagnosis and screening, novel therapies, and significantly improved 5-year survival figures over the past two decades.

Why is the outcome so much worse for gastric carcinoma compared to, say, the other common gastro-intestinal tumour, colonic carcinoma? Three potential reasons come to mind. First, carcinoma of the stomach tends to present at a later stage (Fuchs *et al.* 1995). Second, effective adjuvant or neo-adjuvant therapies are lacking. Third, since the stage-for-stage survival is much lower in gastric carcinoma, it is likely that there is something intrinsically different about its tumour biology. Part of this may relate to the anatomy of the stomach, which has more abundant lymphatic vessels than the colon, especially in the submucosa (Lewin *et al.* 1996), which may predispose to the major problem of loco-regional recurrence (Roukos 2000). For elucidating further aspects of the tumour biology, other means are required; and over the next decades, it is to be hoped that greater understanding of the molecular and cellular biology of cancers in general, and, for our purposes, of gastric carcinoma in particular, will lead to advances in earlier diagnosis and therapy akin to those already seen with other tumour types.

1.5 Classical candidate gene approaches to the molecular and cellular biology of cancer

Histopathogical assessment of cancer has a history stretching back over 150 years but analysis of the many molecular changes occurring in cancers and during their development has become a viable option only relatively recently, although it has long been recognised that cancer is a genetic disease.

This research project uses relatively new large-scale profiling technologies to study the genes expressed in normal stomach and gastric carcinoma. Before moving to a discussion of such methods, however, it is important to review our current understanding of the molecular events underlying cancer. Their description could and does take up entire textbooks rather than an introductory chapter, but a brief summary follows (reviewed in (Strachan *et al.* 1999; Hanahan *et al.* 2000)). It has really been only in relatively recent years that this knowledge has begun to translate to the clinic and impact on the diagnosis (reviewed in (Sidransky 1997)) and treatment of cancer, but these advances are highlighted.

The central dogma of molecular biology is that the genomic DNA is transcribed to RNA which in turn is translated to protein which then acts again on DNA. Except for lymphocytes, normal somatic cells in any one individual have identical DNA, which in humans is now estimated to contain around 30,000-40,000 (protein-coding) genes (Lander *et al.* 2001), of which only a minority are active in any cell at one time. Active genes are expressed as messenger RNA (mRNA) transcripts, which are then translated to produce proteins, which in general perform the function of the gene. Cells from different tissues are therefore characterised by their different patterns of gene activity and hence by differential expression of mRNA and protein. Genes may therefore be studied at any of these three levels: DNA, RNA or protein. First, DNA will be considered: its study has formed the basis for most of the classical work on cancer and cancer-associated genes, as described in the next sections.

1.5.1 Cancer at the level of the genome

Organisms which show hereditary variability evolve by natural selection, and DNA genotypes which confer a reproductive advantage will be selected and will come to dominate the population. Within an organism the same is true of cells: those which derive enhanced proliferative capacity from somatic mutation will become dominant, in the form of a cancer. Protective mechanisms exist to prevent this occurring within organisms, at least until reproduction has taken place, so that when somatic mutations occur they are either repaired or they induce the cell to die.

However, no single DNA mutation is sufficient to transform a normal cell into a cancerous cell. It has been calculated that on average six independent mutations are required, which under normal circumstances would be exceedingly rare: this has been described as the multistep evolution of cancer.

There are two general mechanisms which may make the development, and persistence, of successive mutations more likely. First, some mutations promote cell proliferation and create an expanded target population of cells in which the next mutations can occur: the genes affected are the oncogenes, discussed in Section 1.5.4, which have have been described as "gatekeepers". Obviously, diseases such as chronic gastritis, which promote increased epithelial proliferation and repair (Correa 1992), exert a similar effect, explaining their role in the carcinogenic pathway. Second, some mutations alter the stability of the whole genome, at either the overall chromosomal or individual nucleotide levels, thus increasing the overall mutation rate: the genes affected have been described as "caretakers". Their effects are discussed in Sections 1.5.3 and 1.5.4. Accumulating these successive mutations still takes time so it is no surprise that most common cancers, if sporadic, including gastric carcinoma, are diseases of post-reproductive life.

1.5.2 Classical cancer-associated genes

Although normal somatic cells in an individual contain identical DNA, cells which are either already malignant or are undergoing carcinogenesis contain abnormal DNA, with changes which may be either large-scale, at a chromosomal level, or smaller-scale, at the level of DNA mutations (Strachan *et al.* 1999; Hanahan *et al.* 2000). In cancer, these alterations tend to target certain classical cancer-associated genes, which fall into two main groups, the oncogenes and the tumour suppressor genes. Other important genes include those involved in cell adhesion (e.g. E-cadherin), telomere maintenance (e.g. telomerase), invasion and metastasis, and other processes involved in carcinogenesis.

Oncogenes are genes which tend to promote cell proliferation. The non-mutated form of the oncogene present in normal cells is called the proto-oncogene. Oncogenes can be thought of in five broad classes: secreted growth factors; cell surface receptors e.g. c-erbB2; signal transducers e.g. ras and abl; nuclear DNA-binding proteins e.g. jun; and cell cycle regulators e.g. cyclins. Cellular proto-oncogenes may become activated to function as pro-carcinogenic oncogenes through either quantitative or qualitative mechanisms. For example, c-erbB2 is activated in breast cancer by gene amplification which results in the presence of numerous copies of the gene at DNA levels and therefore increased amounts of functional protein (Eisenhauer 2001). On the other hand, abl is activated in chronic myeloid leukaemia (CML) by a translocation which joins abl on chromosome 9 to a gene called BCR (breakpoint cluster region) on chromosome 22. The resulting cytogenetic abnormality is characteristic of CML and is called the Philadelphia chromosome. The resulting chimaeric BCR/ABL gene produces a tyrosine kinase signal transduction protein which is abnormally truncated, being constitutionally active rather than regulated and hence having a pro-carcinogenic effect (Mauro *et al.* 2002). Both c-erbB2 and BCR/ABL represent ideal therapeutic targets (see later Section 1.5.6). The ras oncogene is similarly activated through a qualitative change, but in this case it is through smaller-scale mutations resulting in amino acid substitutions, usually at codons 12, 13 and 61. Activation of an oncogene is usually a genetically dominant event in that a single mutant allele causing so-called gain of function may alter the phenotype of the cell.

Oncogenes have classically been identified in two ways (Strachan *et al.* 1999). The first was through animal tumour viruses. Various animal cancers were known to be caused by viruses. Investigation of these viruses showed that certain retroviruses contained genes additional to the normal retroviral genome. The extra genes in the retroviruses were found to be responsible for their cancer-causing properties and were thus termed oncogenes. The second method of identifying oncogenes involved cell transformation assays. Mortal animal cells grown *in vitro* can be induced by carcinogenic agents and genetic changes to acquire a different, so-called transformed, phenotype, with invasive behaviour and unrestrained growth potential. NIH3T3 cells in particular readily undergo transformation *in vitro*, usually requiring only one further genetic event. Transfection of these cells with fragments of DNA from cancer cells results in transformation. The transformant cells can be selected and the transforming DNA analysed. Such assays pinpointed essentially the same set of oncogenes as were discovered through viral studies. Soon it was recognised that oncogenes had counterparts in normal animal cells, and these were called the proto-oncogenes, as described previously, with important roles in the regulation of cellular growth and proliferation.

In contrast to the oncogenes, tumour suppressor genes normally tend to inhibit events leading towards cancer (Strachan *et al.* 1999). In cancer cells, the mutant tumour suppressors have lost their function. The protein products of tumour suppressor genes have a variety of normal functions: many are involved in the cellular response to DNA damage, by preventing cell cycle progression or provoking apoptosis; others maintain genome stability and inhibit the acquisition and perpetuation of mutations. Loss of function of a tumour suppressor gene in cancers is a genetically recessive event in that both alleles must be inactivated to alter the phenotype of the cell.

Tumour suppressor genes have classically been identified in two ways (Strachan *et al.* 1999). The first is through positional cloning of genes causing rare familial cancers. Of this one of the best examples is retinoblastoma, which is a rare childhood tumour of the retina. In 40% of cases the tumour is familial and often occurs bilaterally. The standard approach to finding candidate cancer-associated genes involves narrowing the physical size of the candidate intervals using techniques such as meiotic recombination or marker disequilibrium in affected families. The trait was thus mapped to chromosome 13q14, at a locus designated RB1. In these families, somatic cells were heterozygous for markers around 13q14, but the tumours were homozygous: they carried a germ-line mutation in one allele at 13q14 and the corresponding normal allele was then lost in the tumours, fulfilling an earlier hypothesis termed the "two-hit" theory of tumour suppressor genes which was proposed by Knudsen. A more recent adjunct approach is to use expression patterns to narrow the region, i.e. to prioritise for analysis the subset of genes that map to the minimal search interval and are expressed in the involved tissue (Brent 2000).

The second classical method of discovery of tumour suppressor genes is through defining chromosomal locations commonly deleted in tumour cells, by cytogenetic methods such as chromosome banding or comparative genomic hybridisation (CGH) or by the molecular genetic method of analysis of loss of heterozygosity (LOH) (Strachan *et al.*

1999). LOH involves screening paired blood and tumour samples by polymerase chain reaction (PCR) for markers spaced fairly closely across the genome. Candidate locations for tumour suppressor genes are identified on the basis of loss or significantly decreased intensity of a PCR product band on a gel in the tumour sample compared with the normal. Because most tumours show significant genomic instability, there is a relatively high background of non-specific changes so many tumours must be screened to identify the desired specific changes. The technique of CGH was developed to survey DNA copy-number variations across a whole genome, and it enables the identification of large chromosomal segments of deletion in tumours compared to normal DNA. Obviously, areas of amplification, usually associated with oncogenes, can also be identified. The mapping resolution of conventional CGH is limited to around 20 Mb, which has been increased with newer array-based techniques (see Section 1.6.2). CGH results can be confirmed using the technique of fluorescence *in situ* hybridisation (FISH), which has a higher resolution (see Section 5.1.2.2).

Thus, the classical molecular alterations associated with cancer include large-scale chromosomal alterations and small-scale, but multiple, mutations in individual tumour suppressor genes and oncogenes. Not surprisingly, gastric adenocarcinomas show most of these changes, as described the next Sections 1.5.3 and 1.5.4, and some differ according to the histological subtype.

1.5.3 Chromosomal alterations in gastric carcinoma

There is strong evidence for a genetic predisposition to gastric carcinoma, both from epidemiological studies and from case reports of gastric cancer families. Napoleon Bonaparte is perhaps the best-known example of the latter: he, his father, his grandfather, four sisters and a brother all died of gastric cancer, some at an early age (Bevan *et al.* 1999) (although sadly the histological type is not reported!). Overall, around 10% of gastric cancer cases show familial clustering.

First, chromosomal abnormalities will be considered. Their occurrence in gastric carcinoma has been reviewed in (Peddanna *et al.* 1995; Menke-Pluymers *et al.* 1996; Tahara *et al.* 1996; Blok *et al.* 1997; Bevan *et al.* 1999). Chromosomal instability (CI) at a large-scale level is common generally in adenocarcinomas: it involves many losses, gains and rearrangements of chromosomes and usually results in polyploidy or aneuploidy. Accordingly, complex karyotypes, studied by classical cytogenetics and flow cytometry, are common in gastric adenocarcinomas, especially in tumours of advanced stage, although the reported percentage of aneuploid cells varies widely from 9% to 68% (Peddanna *et al.*

1995). Similarly, at a smaller-scale level, cells from typical advanced adenocarcinomas of breast or colon show LOH at around one quarter of all loci. Most of these changes are simply a reflection of general chromosomal instability, and only a few appear to be causally connected with the cancer.

However, certain numerical and structural changes at the large-scale chromosomal level are apparently non-random. In gastric carcinoma, these include: trisomy of 8, 9, 12, 19; loss of the Y chromosome; gain of an X chromosome; chromosome 3 translocations; and further non-random abnormalities of chromosomes 1, 2, 3, 6, 7, 19 and 20. Allelic loss (LOH) has been described for chromosomes 1p, 1q, 3p, 3q, 5q , 7p, 11p, 11q, 12p, 12q, 13q, 17p and 18q (Peddanna *et al.* 1995; Menke-Pluymers *et al.* 1996; Tahara *et al.* 1996; Blok *et al.* 1997; Bevan *et al.* 1999). In some cases, likely targets of the chromosomal aberrations are obvious. For example, 17p contains the tumour suppressor gene p53 while 5q harbours the tumour suppressors APC (adenomatous polyposis coli), MCC (mutated in colon cancer) and DCC (deleted in colon cancer) (see the next Section, 1.5.4). More particularly for carcinoma of the stomach, LOH at chromosome 7q at the locus D7S95 has been associated with a worse prognosis and high incidence of peritoneal dissemination: this site may therefore contain a candidate tumour suppressor gene for gastric cancer (Tahara *et al.* 1996).

In most cases, however, the diagnostic and therapeutic implications of these chromosomal changes remain unclear. Such aberrations are, however, likely to be of clinical relevance in the future, as is now becoming clear for colon carcinoma (see Section 1.7.2.2). In future research efforts, the selection of candidate disease genes may be facilitated by integrating knowledge of the chromosomal regions implicated in gastric carcinogenesis with gene expression profiles and transcript map locations (Brent 2000; Emmert-Buck *et al.* 2000). Candidate genes which have already been implicated in gastric carcinoma are discussed next.

1.5.4 Classical cancer-associated genes in gastric carcinoma

At the level of individual genes, most of the classic cancer-associated molecular alterations have been found in gastric carcinomas. There have been a number of recent reviews, many authored by Eiichi Tahara (Tahara *et al.* 1994; Peddanna *et al.* 1995; Tahara *et al.* 1996; Blok *et al.* 1997; Chan *et al.* 1999; Yasui *et al.* 1999). A summary is presented in Table 1.3. In general, genetic instability, inactivation of tumour suppressors, alteration of cell cycle control and apoptosis genes and telomerase activation are implicated in early gastric carcinogenesis, while over-expression or other activation of oncogenes is linked with cancer progression (Tahara *et al.* 1996). These carcinogenic mechanisms will now be considered individually.

As described above, many gastric carcinomas show large-scale chromosomal abnormalities, known as chromosomal instability (CI), which often cause aneuploidy. Some tumours, however, have a normal, diploid chromosome complement by karyotypic analysis, but exhibit smaller-scale genetic changes because their DNA repair apparatus is faulty and allows errors to develop during DNA replication. When such base-pair mismatches are introduced, the length of marker DNA segments, called microsatellites, changes, which can be used as an assay. These tumours are therefore described as showing "microsatellite instability" (MI) or as being "replication error" positive (RER+). This is characteristic of tumours arising in patients with Hereditary Non-Polyposis Colorectal Cancer (HNPCC or Lynch syndrome) (Peddanna *et al.* 1995). While HNPCC is associated primarily with an increased rick of colonic carcinoma, it also predisposes to carcinomas of the stomach, pancreas and ovary.

The genes altered by MI are different from those targeted in the CI group (which include classical tumour suppressors, see last Section, 1.5.3). Common targets in MI instead include, for example, the receptor for transforming growth factor beta (TGF-β), which is discussed later. Not surprisingly, the germ-line mutations responsible for MI involve genes responsible for DNA mismatch repair, such as MLH1 and MSH2. As shown in Table 1.3, MI is more common in gastric carcinomas of intestinal rather than diffuse type. In addition, MI is an early event in the multi-step process of carcinogenesis, since it is also present in dysplasia (i.e. in adenomas) and in intestinal metaplasia (Tahara *et al.* 1996), both of which are implicated in the pathway leading to intestinal type carcinomas (Correa 1992).

Mutation and allelic loss (LOH) of the tumour suppressor genes p53 and APC are also common in gastric carcinomas of intestinal type, and in their precursors, IM and dysplasia (Tahara *et al.* 1996). p53 protein is responsible for the cell cycle arrest of cells with damaged DNA. In normal cells, p53 is expressed at very low levels and is rapidly degraded, but it is activated by stresses such as irradiation or heat. Amongst its many functions, p53 then activates a number of genes including the cyclin-dependent kinase (CDK) inhibitor p21$^{(\mathrm{WAF1/CIP1})}$, which binds to and inhibits G1 cyclin-CDK complexes. This causes cell cycle arrest in G1 phase, preventing progression to S phase, until the damaged DNA is repaired. p53 can also provoke apoptosis. Given these important functions (such that p53 has been called the "guardian of the genome"), it is not surprising that mutations in p53 are the most common of all genetic changes in human cancers, with inactivation being caused mainly by mutations in the highly conserved region spanning exons 5-8.

Around 70% of gastric carcinomas harbour p53 mutations, although the spectrum of p53 mutations differs between gastric carcinoma and oesophageal tumours, as well as between intestinal and diffuse gastric carcinomas, implying different underlying carcinogenic mechanisms (Peddanna *et al.* 1995; Blok *et al.* 1997).

Unlike p53, the tumour suppressor gene adenomatous polyposis coli (APC) is altered only in intestinal-type gastric carcinoma and its precursor conditions. Amongst various functions, normal APC inhibits the ability of the Wnt protein to activate expression of the c-myc gene. In the absence of normal APC, c-myc is inappropriately activated and induces transcription of many genes needed for cell cycle progression from G1 to S phase, and therefore enhances cell proliferation. APC also normally regulates the cytoplasmic protein β-catenin by binding to it. Mutated APC lacks this function, allowing β-catenin to accumulate and enter the nucleus where it again affects gene transcription. APC's name indicates its importance in colorectal cancer, which may explain its frequent alteration in intestinal, but not diffuse, gastric carcinomas.

Instead, at least 50% of diffuse gastric carcinomas lack the function of another gene which functions as a tumour suppressor. E-cadherin is a calcium-dependent transmembrane adhesion molecule responsible for cell-cell connections in epithelial cells. It is important in establishing cell polarity and maintaining normal tissue morphology and differentiation. Normally, inside the cell, cadherins are linked by the previously mentioned catenin molecules to the actin cytoskeleton: β-catenin thus provides a link with APC. Inactivation of E-cadherin or β-catenin is an early event in the development of diffuse gastric cancer, and may in fact occur as a germ-line mutation, as discussed in the next Section, 1.5.5. In addition to APC, the E-cadherin/β-catenin complex is functionally linked with proto-oncogene tyrosine kinase receptors including c-erbB2/neu and c-met (Tahara *et al.* 1996).

And indeed, gastric carcinomas exhibit alterations in the structure and function of proto-oncogenes. The aforementioned c-met encodes a transmembrane tyrosine kinase which is the receptor for hepatocyte growth factor/scatter factor (HGF). HGF is a growth factor with mitogenic and motogenic properties as well as the ability to induce epithelial cell invasion in collagen matrices *in vitro*, and a likely role in tumour metastasis *in vivo* (Blok *et al.* 1997). c-met activation can be due to: gene amplification with subsequent protein over-expression; defective post-translational processing of the precursor protein; or oncogenic rearrangement between two genetic loci, met and tpr, leading to a novel 5.0 kb mRNA transcript encoding a 65 kDa fusion protein (Blok *et al.* 1997). c-met is overexpressed in up to 40% of gastric carcinomas, more commonly in but not limited to those of diffuse type (Tahara *et al.* 1996).

A further interesting aspect of c-met/HGF and associated adhesion molecules is their role in cancer-stromal interactions, which may contribute to the morphogenesis of the two different histological subtypes of gastric cancer (Tahara *et al.* 1996). Stromal cells around tumours secrete HGF when activated by growth factors or interleukins. When gastric carcinoma cells express E-cadherin and catenin, HGF promotes the formation of glandular structures and hence an intestinal phenotype. Conversely, where cell adhesion molecules are lost, then HGF tends to cause scattering of tumour cells and thus a tumour of diffuse morphology (Tahara *et al.* 1996).

Other proto-oncogene transmembrane tyrosine kinase receptors are also implicated in gastric carcinogenesis, again with molecular differences echoing the histology. k-sam encodes the receptor for keratinocyte growth factor and is amplified only in diffuse cancers (Tahara *et al.* 1996). Amplification of c-erbB2, however, is detected only in gastric carcinomas of intestinal type (Tahara *et al.* 1996), and likewise is common in colonic carcinomas. Therapeutic advances involving c-erbB2 in breast cancer in particular are discussed later in Section 1.5.6.

As well as their receptors, growth factors (GF) themselves play a role in cancer of the stomach. Epidermal growth factor (EGF) and related peptides, including transforming growth factor alpha (TGFα) and cripto (reviewed in (Yarden 2001)) are common positively-acting growth factors for both types of gastric carcinoma, in which they are often over-expressed in the absence of underlying gene amplification. Some GFs, such as cripto, are also over-expressed in the precursor condition of IM.

TGFβ, platelet-derived growth factor (PDGF) and basic fibroblast growth factor (bFGF) are all commonly over-expressed in diffuse gastric cancers (Tahara *et al.* 1996). TGFβ is a potent growth inhibitor, with tumour-suppressing activity. Cancers in general often fail to respond to this growth inhibition either because of genetic loss of TGFβ signaling components (such as occurs in MI tumours, described earlier) or, more commonly, because of disturbances of the signaling pathway downstream, for example through ras activation. Carcinomas often secrete excess TGFβ but respond to it abnormally by enhanced invasion and metastasis. Therapeutic approaches should thus aim to inhibit the TGFβ-induced invasive phenotype, but simultaneously to retain its growth-inhibitory and apoptosis-inducing effects (Akhurst *et al.* 2001).

The ras proto-oncogene family, mentioned above, includes the homologous H-, K- and N-ras genes, which code for closely related 21 kDa proteins ($p21^{RAS}$). The ras genes are activated by point mutation, especially at codons 12, 13 and 61. K-ras mutations are frequent in both colonic and pancreatic adenocarcinomas, but its role in gastric carcinoma appears to be less important where, not surprisingly, mutations are limited to tumours of intestinal type but occur in only around 9-18% (Tahara *et al.* 1996; Blok *et al.* 1997).

Cyclins, cyclin-dependent kinases (CDK) and their inhibitors regulate the cell cycle and hence cell growth, survival and death (Tahara *et al.* 1996). As with the tyrosine kinase receptors, it seems that different cyclins and related molecules play a role in different cancers: over-expression of cyclin D1 is common in oesophageal squamous carcinoma but not in gastric and colorectal carcinomas, which instead exhibit amplification of cyclin E, which may be associated with an increased risk of lymph node metastasis. p21$^{(WAF1/CIP1)}$ inhibits cyclin/cdk2 kinase and is, as previously discussed, induced by wild-type p53. p27$^{(KIP1)}$ also inhibits cyclin/CDK complexes, and both act as negative regulators of cell cycle progression. Overexpression of the bcl-2 molecule inhibits apoptosis and is more common in intestinal tumours.

Telomerase activation is also important in the development of gastric carcinoma. The telomeres are structures which lie at and protect the ends of chromosomes. In normal somatic cells, the telomeres shorten progressively at each cell division, a process which eventually leads to cell senescence. Telomerase stabilises the telomeres and confers immortality (Tahara *et al.* 1996). Telomerase has two components: the RNA component forms a template on which the other, protein, component acts, as a reverse transcriptase. Telomerase activity is absent from most normal somatic cells but is restored in most tumours. Both the protein and RNA components of telomerase are present in gastric carcinomas, of both types, as well as in the precursor IM. As with many of these molecular changes, telomerase activation has been found to parallel morphological abnormalities during cancer progression. The nuclei of the dysplastic epithelial cells in gastric adenomas differ according to whether telomerase activity is strong, weak or absent. Strong telomerase activity is associated with large irregular nuclei, with a round or oval rather than (close to normal) spindle shape, and with coarse chromatin and prominent nucleoli: these nuclear features indicate progression towards malignancy (Yasui *et al.* 1999).

Moving then to genes involved in metastasis, CD44 is, like E-cadherin, important in cell-to-cell adhesion. Variant CD44 proteins due to alternative splicing of exons 6-14, containing the extracellular domain, are expressed in many cancers. Abnormal CD44 transcripts can be detected in most gastric carcinomas but their types differ between intestinal and diffuse cancers (Tahara *et al.* 1996). CD44 variants containing the intron 9 sequence are expressed prefentially in both gastric and colorectal carcinomas suggesting a common role in gastro-intestinal carcinogenesis (Tahara *et al.* 1996). nm23 is a candidate metastasis suppressor gene which encodes a nucleotide diphosphate kinase and transcription factor for c-myc (Tahara *et al.* 1996). LOH at the nm23 locus is found in 8% of primary gastric tumours, but is not surprisingly more common in metastatic disease. The latter is also associated with increased levels of urokinase-type plasminogen activator and plasminogen activator inhibitor in tumour tissue (Tahara *et al.* 1996).

Gastric carcinomas thus contain multiple genetic abnormalities. Some are common events in both histological subtypes, including: genetic instability, loss of function of p53, telomerase activation, amplification of c-met and cyclin E, and variant CD44 transcripts. Other molecular changes differ according to the tumour morphology, providing further evidence for the existence of different underlying carcinogenic mechanisms. Diffuse cancers show reduced expression of cell adhesion molecules, such as E-cadherin, and amplification of K-sam. These aberrations are not found in gastric carcinomas of intestinal type which instead show loss of function of APC, amplification of c-erbB2 and mutations of K-ras, changes which are also seen in the precursor conditions of intestinal metaplasia and dysplasia (Tahara *et al.* 1994; Tahara *et al.* 1996), lending further support to Correa's proposed pathway for the development of distal gastric carcinoma. Furthermore, the molecular changes seen in gastric carcinomas of intestinal type resemble those in colonic carcinomas, in which there is clear evidence for an adenoma-carcinoma sequence with associated genetic changes (Fearon *et al.* 1990). In fact, a similar "Vogelstein-o-gram" of the molecular changes accompanying the histological steps in Correa's pathway has been proposed for distal gastric carcinogenesis, but is not yet generally accepted. The similarities between gastric and colonic adenocarcinomas suggest that they may share some pathogenetic factors.

While these molecular investigations have greatly elucidated the biology of gastric carcinoma, and many of these genetic changes correlate with tumour behaviour, none as yet has translated to the clinic for use as an independent diagnostic, prognostic or therpeutic marker, with the notable exception of E-cadherin, as discussed in the next section. Gastric carcinoma is thus a heterogeneous disease with varying epidemiology, risk factors, pathogenesis and molecular pathology depending on its site and histology. Perhaps partly because of this complexity, stomach cancer has been studied in less detail than colorectal and other carcinomas, but merits further research due to its clinical importance and our emerging understanding of its causes.

1.5.5 Recent advances in the molecular genetics of gastric carcinoma

Our expanding knowledge of the molecular events underlying cancer has only recently begun to translate to the clinic and to make an impact on the diagnosis and treatment of cancer. Advances have as yet been limited in gastric carcinoma, but one success story involving a classical cancer-associated gene is that of E-cadherin. As described in the previous section, E-cadherin is a calcium-dependent adhesion molecule responsible for cell-cell connections in epithelial cells. Mutations in E-cadherin are the first to have been found in familial gastric cancer (Guilford *et al.* 1998). Germ-line truncating mutations in the E-cadherin gene have been found in families with hereditary diffuse gastric cancer, which has an autosomal dominant pattern of inheritance and occurs mainly in young persons. This was first identified in kindreds from New Zealand but has since been confirmed elsewhere in the world.

In a recent study, five carriers of mutations, who were between 22 and 40 years old, underwent total gastrectomy prophylactically (that is, they were then asymptomatic but were expected to have sub-clinical disease or to develop disease later). In each case, the mucosa of the stomach was extensively sampled for histopathology and in each case superficial infiltrates of the malignant signet-ring cells of diffuse gastric carcinoma were identified. Since the tumours were superficial and thus at an early stage, the prognosis was good. For families with a strong history of diffuse gastric carcinoma, therefore, genetic counselling, E-cadherin genetic analysis and consideration for prophylactic gastrectomy in mutation carriers are now recommended (Huntsman *et al.* 2001). Not surprisingly, somatic inactivation of E-cadherin through mutation (of which over 70% are complete or partial deletions of exons) and LOH is important in the pathogenesis of sporadic diffuse gastric carcinomas, as well as in lobular carcinoma of the breast. Loss of E-cadherin function could enable cells to escape from cell-to-cell growth-control signals. Alternatively it could increase the availability of free cytoplasmic β-catenin, in a manner akin to that of that other tumour suppressor gene APC.

One of the first genetic factors, except for blood group A, identified as predisposing to gastric adenocarcinoma of intestinal type, is polymorphisms in the interleukin-1 gene. Interleukin-1 is secreted by lymphocytes as part of the immune response in the gastric mucosa to *H. pylori* infection. Certain allelic variations, or mutations, in the interleukin-1 gene predispose to the development of body-predominant gastritis, atrophy and hypochlorhydria, and gastric carcinoma, in response to infection with *H. pylori* in susceptible individuals (El-Omar *et al.* 2000). These interleukin-1 polymorphisms lack any known effect in the absence of gastric *H. pylori* infection. This discovery emphasises the fact that therapy to eradicate *H. pylori* is definitely to be recommended in families with a high incidence of (intestinal, distal) gastric carcinoma.

1.5.6 The impact of classical cancer-associated genes on clinical practice

While molecular medicine is only now beginning to make an impact in the management of gastric carcinoma, in other tumour types there have already been major successes in cancer prognostics and therapeutics.

One of the oldest molecular tests with clinical significance is the assessment of n-myc amplification in neuroblastoma, which is an aggressive tumour of childhood. N-myc is amplified in a third of primary tumours; these patients have a worse prognosis than those without the gene amplification. Their survival has been improved by treating them more intensively from their initial diagnosis (Brodeur *et al.* 1993).

Further examples of molecular pathology directing therapy are provided in breast cancer. It has long been known that breast carcinomas are hormonally sensitive: indeed, it was George Beatson, who founded the laboratories in which this research was performed, who first used oophorectomy (surgical removal of the ovaries) for its treatment. The major source of this hormone sensitivity is the oestrogen receptor, a nuclear DNA-binding protein, which is expressed by some but not all breast cancers. The presence of oestrogen receptors used to be assayed biochemically in protein extracts of the tumour tissue. It is now routinely demonstrated by immunohistochemical staining of paraffin-embedded sections of breast carcinoma, which are then assessed microscopically by a histopathologist. The presence of oestrogen receptors in the tumour indicates that it is likely to respond to hormonal therapy, and in particular, to tamoxifen (Osborne 1998). Tamoxifen is now one of the most widely used and effective anti-cancer drugs, yet because it is not a general anti-proliferative agent but instead is directed against a target rather specific to the cancer type, it is relatively non-toxic.

Breast cancer treatment has also benefited from a novel agent more recently developed. The cell surface receptor and proto-oncogene c-erbB2 (HER2/neu) has been described previously. Like n-myc, c-erbB2 is activated by gene amplification which results in increased amounts of functional protein (Eisenhauer 2001). The subset of breast cancers (25-30%) which over-expresses c-erbB2 tends to lack oestrogen receptors and therefore does not usually respond to tamoxifen. Because oestrogen receptor positivity is associated with cellular differentiation, the tumours with excess c-erbB2 also tend to be more poorly differentiated and to have a worse prognosis (Eisenhauer 2001). A humanised monoclonal antibody directed against erbB2 has been developed. The antibody binds to the extracellular domain of c-erbB2 and abolishes its signal transduction activity, probably by preventing receptor dimerisation. It is marketed under the name Trastzumab or Herceptin. About 15% of patients with metastatic breast carcinoma who have previously not responded to conventional therapies show an objective response with Herceptin (Eisenhauer 2001).

More recently still, the drug GLIVEC or STI571 has become available (Blume-Jensen *et al.* 2001; Mauro *et al.* 2002). It is a small molecule inhibitor of the tyrosine kinase receptor proto-oncogenes, which are over-expressed in a number of cancers, in particular chronic myelogeneous leukaemia (CML), gastro-intestinal stromal tumours (GISTs) and possibly malignant melanoma. The disease-causing genetic aberration in CML is the Philadephia chromosome, which results in a chimaeric BCR-ABL protein (see Section 1.5.2). Compared with wild-type c-abl, the chimaera has increased tyrosine kinase activity, which is essential for its transforming effect. A tyrosine kinase inhibitor should thus be an effective and selective agent for CML; STI571, which blocks the ATP binding essential to kinase function, is the successful proof-of-principle, and is effective in both chronic phase CML and blast crisis. Amongst other tumours, the activation by mutation of c-kit is likely to be a critical event in the pathogenesis of GISTs, which previously were refractory to therapy but in which STI571 is effective in metastatic disease (Mauro *et al.* 2002). Its potential for use against the aggressive and unpredictable malignant melanoma is also exciting.

Thus, although the lead time is long, improved understanding of the molecular events underlying tumorigenesis can bring about worthwhile developments in terms of better cancer prognostication and treatment, and presumably also prevention and early diagnosis. Identification and investigation of further candidate genes are therefore likely to yield only more improvements in cancer care.

1.6 Large-scale profiling of cancer molecular genetics

The research on classical cancer-associated genes as described above has usually involved identification of the individual gene at the DNA level. Investigations of the expression of these genes at the levels of mRNA and protein were usually secondary to, and dictated by, the initial DNA studies; and all of these usually focussed on a single gene. Using these traditional DNA-based methods, perhaps one to two hundred classical cancer-associated genes have been identified.

The human genome, however, contains many more genes than this. Current estimates based on the genome sequence and on Expressed Sequence Tags (ESTs, see

Section 1.6.1.2) suggest that there are around 30,000-40,000 protein-coding genes (Ewing *et al.* 2000; Lander *et al.* 2001; Venter *et al.* 2001). While single gene studies remain valuable and necessary, new technologies have emerged over the past decade which enable these thousands of genes to be studied simultaneously ((Brent 2000) provides an excellent review of genomic biology). As stated previously, the central dogma of molecular biology is that genomic DNA is transcribed to RNA which in turn is translated to protein which then acts again on DNA. Cells from different tissues are characterised by their different patterns of gene activity and hence by differential expression of mRNA and protein. Of the total of 30,000-40,000 genes, only around 5,000-10,000 genes are expressed in a given cell type (Adams *et al.* 1995; Bains 1996; Zhang, L. *et al.* 1997). It is likely that many of these genes will also play a role in the development and progression of cancers, yet their importance is currently not recognised. Their study on a large-scale basis could theoretically be at the levels of DNA, RNA or protein.

Obviously in cancers, there are DNA changes in genes. The high-throughput study of individual genes on a genome-wide scale is made difficult, though not impossible, by the fact that only about 1.5% of the human genome consists of coding sequence (Lander *et al.* 2001). As will be discussed later in Section 1.6.2, methods addressing this problem have emerged over the past year or two, but initially the other molecules could be more easily studied.

The active coding DNA in any cell is reflected in its RNA expression pattern. Like DNA, RNA is made up of only four nucleotides (Strachan *et al.* 1999). In this respect, RNA may be more easily studied than the resulting proteins, which are built up from 20 amino acids and which are also subject to significant post-translational modification such as phosphorylation and glycosylation. Again, as will be discussed later in Section 1.6.2, methods exist to address this problem, but for now RNA will be our focus.

1.6.1 Gene expression profiling at the mRNA level

Genome-wide surveys of cellular mRNA are now possible through the use of the complete Human Genome sequence and the availability of methods for large-scale expression profiling. But why should this be a worthwhile and popular aim? A recent review article clearly presented the reasons (Brown *et al.* 1999). First, the function of a gene is tightly linked to its expression pattern. Each gene is expressed in the specific cells and under the specific conditions in which its product makes a contribution to the fitness of the cell and organism. Just as natural selection has tuned the biochemical nature of the protein product, so it has tuned the regulatory properties that govern when, where and in what amounts the product is made. Second, promoters function as transducers, responding to data about the external environment and internal state of a cell by changing the transcription of specific genes, so that, through gene expression patterns, the functions of each promoter can be elucidated. Third, the set of genes expressed in a cell determine what the cell is made of and what it can and cannot do. Such expression patterns and knowledge of what the individual genes do provide a biochemical picture of the living cell which is the molecular counterpart of traditional morphology. And lastly, the technologies now available make such experiments feasible (Brown *et al.* 1999).

For cancer research in particular, it has long been recognised that describing unique gene transcription patterns in cancers would provide an alternative approach to single gene studies for the development of rational approaches to early detection and better prognostication and treatment. This is the central tenet of the Cancer Genome Anatomy Project (Kuska 1996; Strausberg *et al.* 2001) which is described in more detail in Section 1.7.

Over the past decade various techniques have been developed to enable and improve the large-scale assessment and comparison of gene expression at RNA level (reviewed in (Carulli *et al.* 1998)). Many of these methods depend on RNA's biochemical characteristics, as follows. Of the total RNA in a cell, only around 1-5% codes for protein (Strachan *et al.* 1999). This is termed messenger RNA (mRNA). It bears a 3' poly(A) tail, consisting of a 200-250 bp length of adenine (A) nucleotides. Such polyadenylation is signalled by the sequence AAUAAA 10-35 nucleotides upstream of the start of the poly(A) tail. In animal cells, all mRNAs bear poly(A) tails except those for histones. The remaining 95-99% of total RNA does not code for protein: the RNAs are thus termed ncRNAs (non-coding) (Eddy 1999). These include the abundant ribosomal, transfer, and small nuclear and nucleolar RNAs, as well as the telomerase RNA component and other species. The poly(A) tail of mRNA provides a convenient method for its separation from the other RNA species, through the use of specific base-pair hybridisation to oligo(dT), in the form of a coating on cellulose, beads or inside tubes.

Oligo(dT) selection is used in all of the methods for studying RNA described in the following sections, except for the first, and original: the Northern blot. After its description, there follow discussions of: expressed sequence tags; reverse-transcription polymerase chain reaction (RT-PCR); subtractive hybridisation and differential display; DNA microarrays; and finally serial analysis of gene expression (SAGE), which was used in this research.

1.6.1.1 Northern blotting

The traditional method of studying RNA is the Northern blot (see Section 2.2.4.2). Purified RNA samples are first size-fractionated by gel electrophoresis, transferred to a membrane then hybridised to a labelled gene or cDNA probe. This permits the assessment of the presence of an RNA species, estimation of its size, approximate determination of its abundance relative to a control transcript and investigation of splice variants. Northern blotting allows simultaneous analysis of many RNA samples (up to around 20 per blot) but with only a limited number of probes (usually one per blot). In some respects, the Northern blot remains the gold standard although RT-PCR, which takes less time and is technically easier, is now often used in its place (see Section 2.2.4.1). However, Northern blotting remains in essence a single gene approach.

1.6.1.2 cDNA libraries and EST sequencing

Expressed sequence tags (ESTs) have been invaluable tools for gene discovery and also for assessing differential gene expression (Adams *et al.* 1991; Adams *et al.* 1995). ESTs are uncharacterised cDNAs (sequences of DNA complementary to mRNA) created through traditional cDNA library technology (see Section 2.2.3.1.2). A cDNA library is created by preparing RNA from a cell or tissue population then reverse transcribing it, usually using an oligo(dT) oligonucleotide primer which binds to and therefore selects the poly(A) tail of the mRNA. These cDNA fragments are then ligated into plasmid vectors and transformed into bacteria. Clones are picked randomly from these libraries and then a single sequencing reaction ("single pass") is performed from a large number of clones. Each sequence provides a unique sequence tag for a particular transcript. In the early 1990's, when this technology began, each sequencing reaction generated around 300 bp or so, but now over 700 bp is more common (Wheeler *et al.* 2001).

Large-scale EST sequencing is labour-intensive and expensive so it is not a technology which can easily be performed in smaller centres. However, large EST databases (dbEST) are available on-line via the web-site of the National Centre for Biotechnology Information (NCBI), along with powerful tools for their analysis and comparison (Wheeler *et al.* 2001). The EST databases now contain over 4 million sequences within hundreds of cDNA libraries from many distinct organs and tissues. These have been mined intensively during this research (see Section 2.2.3.2 amongst many).

1.6.1.3 Reverse-transcription polymerase chain reaction (RT-PCR)

For an RT-PCR experiment, RNA from a given cell or tissue sample is reverse transcribed to cDNA. Then specific oligonucleotide primers are used to amplify the transcript of interest by the polymerase chain reaction (PCR) (see Section 2.2.4.1). The PCR product will be of a certain predictable length and can be visualised by gel electrophoresis as a band. This permits the assessment of the presence of an RNA species. Approximate determination of its abundance relative to a control transcript can be undertaken: such quantitation is made easier by newer techniques using fluorescent probes and automated machines. However, RT-PCR, like Northern blotting, is essentially a single gene approach, although it is faster, much more sensitive, and technically rather simpler.

1.6.1.4 Subtractive hybridisation methods and differential display

These methods were described in the early 1990's and are based on RT-PCR technology. They allow the identification of differentially expressed genes, particularly those showing a relatively high abundance or large expression difference. Because these techniques are comparative, the data produced are not for a single sample, but represent differences between two specimens.

Subtractive hybridisation and cloning methods (SH) have been used for many years but their adaptation for PCR made the method much faster and easier with the ability to use less starting material. In SH, two cDNA samples are compared: a normal "test" cDNA and a deleted "driver" cDNA. The test cDNA is mixed with a large excess of driver cDNA, denatured and re-annealed. By various means, double-stranded cDNAs are selected in which both strands are composed of test cDNA. These preferentially represent sequences in the test cDNA that are absent from the driver cDNA, that is, differentially expressed transcripts (Lee *et al.* 1991). Reverse-transcriptase representational difference analysis (RT-RDA) uses broadly similar methods to enrich for and identify sequences present in the test but not in the driver sample (Hubank *et al.* 1994).

mRNA differential display (DD) is also a PCR-based method which can study the expression of many genes simultaneously, through the use of partically degenerate primers (Liang *et al.* 1992). It uses a modified oligo(dT) primer which has a different single nucleotide or dinucleotide at the 3' end causing it to bind to the poly(A) tail of a subset of mRNAs. For example, if the oligonucleotide TTTTTTTTTTTTGA is used as a primer, it will preferentially prime cDNA synthesis from those mRNAs where the dinucleotide TC precedes the poly(A) tail. The second primer which is used is usually an arbitrary short sequence. This is often ten nucleotides long, but, because of mismatching, especially at the 5' end, it can bind to many more sites than would be theoretically expected for a decamer.

PCR amplification results in a complex ladder of multiple product bands which are resolved by size-fractionation by polyacrylamide gel electrophoresis (PAGE, see Section 2.2.2.7). DD is useful in the large-scale comparison of mRNA populations between different samples and in the cloning of differentially expressed genes. For the latter, specific PCR bands which are present in one source but absent from another are isolated from the gel, amplified further by PCR and sequenced for further analysis.

These techniques are powerful but often technically difficult. There are also other limitations. Because the methods are comparative, the data produced are not comprehensive for a single sample, but represent differences between two specimens. False-positive results may also be a problem, and large differences may be needed for the genes to be identified as differentially expressed.

1.6.1.5 Oligonucleotide or cDNA expression arrays

Although the previously described techniques have proven extremely useful, a more systematic and comprehensive method would be preferable for large-scale expression analyses. One way of achieving this has been through DNA microarrays (Schena *et al.* 1995; Schena *et al.* 1996), which have been reviewed extensively (Brown *et al.* 1999; Lockhart *et al.* 2000).

In essence, microarrays act as reverse Northern dot blots, with probes fixed to a solid surface, to which the test RNA is added. There are two main methods. cDNA arrays are made by spotting cDNAs onto glass slides or another solid support (Brown *et al.* 1999). Oligonucleotide arrays (e.g. Affymetrix) can be made by direct synthesis of oligonucleotides on chips using photolithographic masking techniques from the semiconductor industry. Both forms of microarray can bear thousands of different sequences, and are used similarly. To compare the relative abundance of genes expressed in two tissues, RNA is isolated then the two samples are labelled with different fluorescent dyes, say one red and one green. The labelled samples are then mixed and hybridised onto the arrayed DNA spots. After hybridisation, the DNA spots are illuminated microscopically and their fluorescence assayed. The measurements are used to determine the ratio, and in turn the relative abundance, of the amount of each gene in the two samples. To compare many samples, multiple pair-wise experiments are performed, each with one sample in common as a control, say from pooled cell lines. Lower-density arrays of cDNA have also been created on more traditional membranes.

Microarrays have mostly been used as yet to monitor RNA expression levels, of which some exciting applications are discussed in Section 1.7.2.2, but measurement of variation at the DNA level is also possible with microarrays, as described in Sections 1.6.2

and 1.7.2.2. This technology is hugely powerful and widely applicable but until recently the general academic community, especially outwith the US, has had limited access. Partly for this reason, this project instead used the alternative technology of SAGE.

1.6.1.6 Serial analysis of gene expression (SAGE) in brief

Serial analysis of gene expression (SAGE) is a method with the ability to monitor the expression patterns of thousands of genes simultaneously. It was first described in 1995 (Velculescu *et al.* 1995; Zhang, L *et al.* 1997)).

SAGE is a patented large-scale mRNA profiling technology which produces comprehensive, quantitative and reproducible gene expression profiles. Unlike the alternative technologies of differential display and subtractive hybridisation, SAGE produces a full catalogue of all transcripts, not only (a subset of) differentially expressed genes; and unlike smaller arrays, SAGE needs no assumptions about the genes which are likely to be expressed, thus allowing the identification of novel genes ((Madden *et al.* 2000; Velculescu *et al.* 2000; Polyak *et al.* 2001) are amongst many excellent reviews).

SAGE is based on generating clones of concatenated (linked) short sequence tags derived from mRNA from the target cells or tissue. Each tag is 9 or 10 bp long and represents one mRNA; and each clone insert contains up to 40 tags joined serially. Sequencing of multiple concatenates therefore describes the pattern and abundance of mRNA, with an improvement in efficiency of up to 40-fold compared with conventional analysis of expressed sequence tags (ESTs). The mRNA transcript corresponding to the short SAGE tag is identified from genetic databases using appropriate software. Because SAGE is labour-intensive and hence limited to small numbers of specimens, the resulting candidate genes are usually validated in a larger set of samples (Velculescu *et al.* 1995; Zhang, L. *et al.* 1997; Velculescu *et al.* 2000). With SAGE, thousands of transcripts can be analysed simultaneously, such that evaluation of most or all of the estimated 10,000 genes expressed in a given cell population is a realistic goal; and SAGE also facilitates the detection and identification of novel genes, often difficult with earlier technologies. An on-line public SAGE database, SAGEmap, has now been made available through CGAP and NCBI (Wheeler *et al.* 2001). SAGEmap now contains over 4 million tags, a figure similar to the number of ESTs in dbEST. A more detailed description of the SAGE method and its advantages and challanges is presented in Chapter 3.

All of these methods for expression profiling are useful in their appropriate context. They vary in the amount of input mRNA required, in the specimen throughput possible and in their sequencing and bioinformatics requirements (reviewed in (Carulli *et al.* 1998)).

The choice of technique is thus dictated by the experimental question(s) being addressed, the technical capability of the laboratory and the biological samples available.

1.6.2 Large-scale profiling: RNA, DNA, protein or all three?

These large-scale profiling technologies for investigating gene expression at the levels of mRNA are hugely powerful, and it is remarkable to note that none were described more than a decade ago. The pattern and levels of mRNAs reflect transcriptional activity, splice variants and mRNA stability and degradation. mRNA levels are immensely informative about cell state and activity of genes, and for most genes, changes in mRNA abundance are related to changes in protein abundance (Gygi *et al.* 1999; Pradet-Balade *et al.* 2001). Yet mRNA is not the whole story.

If we first stay with gene expression, obviously proteins are the final downstream product of the mRNA intermediate. The new methods of proteomics enable their study in a large-scale manner (Pandey *et al.* 2000). Proteomics means the large-scale analysis of proteins. The proteome is the protein complement of a cell. The word is traditionally associated with two-dimensional gel polyacrylamide electrophoresis of protein extracts from different samples. This yields a display of a large number of proteins. Such gels have been used since the 1970's but reproducible electrophoresis and image analysis for their study and comparison have become available only more recently. However, the most significant breakthrough in proteomics has been the mass spectrophotometric identification of gel-separated proteins. The gel spots of interest are cut out, the protein extracted, and subjected to sequencing. This relies on the digestion of gel-separated proteins into peptides by a sequence-specific protease such as trypsin. The pattern and levels of proteins reflects translational activity, post-translational modifications such as phosphorylation and glycosylation, and protein stability and degradation. However, protein-based approaches are not without problems. They are generally more difficult, less sensitive and have a lower throughput than mRNA methods. Even the best two-dimensional gels can routinely resolve no more than 1,000 proteins and abundant proteins in samples, especially serum, can overshadow the rest. Nevertheless it is a powerful method, as seen from the clinical example given later in Section 1.7.2.2.

Moving back to DNA, large-scale array profiling methods have been developed for the analysis of DNA changes at both chromosomal and nucleotide levels. The identification of genomic imbalances (either gains or losses in copy number) by comparative genomic hybridisation (CGH) (see previous Section 1.5.1) can now be performed on a high-density array, using either large genomic clones (Pinkel *et al.* 1998) or cDNA sequences from over 30,000 radiation-hybrid mapped human genes (Pollack *et al.* 1999). At the nucleotide level, genome-wide profiling is becoming possible through the use of single nucleotide polymorphisms (SNPs), which are single nucleotide positions in the genome sequence for which two or more alternative alleles are present at appreciable frequencies, traditionally at least 1%, in the human population (Strachan *et al.* 1999). SNP chips enable the study of LOH and allelic imbalance between normal and tumour samples, as seen from the clinical example given later in Section 1.7.2.2. SNP chips will also facilitate genome-wide linkage disequilibrium mapping of human disease genes (Wang *et al.* 1998; Lander *et al.* 2001).

Continuing at the DNA level, rather than scanning the genome, it is possible instead to focus in on certain genes. For example, the previously discussed tumour suppressor gene p53 is the most frequently mutated gene in human cancer. 30-70% of tumours bear a point mutation in one of the two p53 gene copies and have lost the other allele, and these changes are usually associated with a worse prognosis clinically. The gold-standard method for their detection is direct sequencing but this is laborious and time-consuming. Arrays bearing all of the mutations identified in p53 can now be substituted (Ahrendt *et al.* 1999).

Ideally, methods for studying DNA, mRNA and protein would be used in parallel, because each provides different but complementary, and sometimes overlapping, information. None of DNA, mRNA or protein is intrinsically more suited to study, and the choice relates instead to the topic under study and the research question being addressed.

1.7 Applications of large-scale profiling of cancer molecular genetics

1.7.1 Cancer Genome Anatomy Project (CGAP)

At the start of this section, it was suggested that describing unique gene transcription patterns in cancers should lead to the development of rational approaches to early diagnosis, prognostication and therapy as well as shedding light on cancer biology. Thereafter, the methods available for large-scale molecular genetic profiling were described. Before stating the aims of this project on gene expression in normal stomach and gastric carcinomas, it would be worth discussing what has already been achieved using these technologies. First, a large governmental project, CGAP, is described. Thereafter the focus is on recent exciting and clinically significant research.

One of the early, and still one of the major, efforts in this area is the Cancer Genome Anatomy Project (CGAP) of the National Cancer Institute in the US (Kuska 1996;

Strausberg *et al.* 2001). Its overall aim is to achieve a comprehensive molecular characterisation of normal, precancerous and malignant cells. As its publicity has stated, "To peer into a single cell and read its molecular signature will enable us to identify precisely what is different between a normal cell and a cancer cell and to follow the genetic changes that take a cell from normality to cancer. It will uncover specific and sensitive molecular markers for cancer detection and will identify molecular differences between tumours, even those that appear to be identical by histological analysis. It will allow us to determine, at the earliest possible stage of cancer development, those tumours that will respond to therapy, which therapies they will respond to and whether a particular cancer will grow quickly or slowly and whether it will metastasise or not." This could be the mission statement of much of the recent research described in the next section, and also of this project.

CGAP's specific objectives included the production and sequencing of cDNA libraries from what it regarded as five of the most important human cancers from the prostate, ovary, breast, lung and colon. (Note that gastric carcinoma is omitted from this "big five" although it is the fourth most common cause of cancer death worldwide.) This then allowed the development of a publicly available human UniGene set of most, if not all, genes that are expressed in tumours (Schuler 1997), and other on-line analysis tools that facilitate the *in silico* analysis of the large CGAP datasets (Wheeler *et al.* 2001). CGAP has been hugely successful in its goal of creating a free interface between genomics and cancer research, and has been heavily used in this project (Strausberg *et al.* 2001).

1.7.2 Recent discoveries made using large-scale expression profiling

Most of the work described here has used either SAGE or cDNA or oligonucleotide microarrays, and all of it has been published within the past few years (mostly this year or last), after this project was started. Genomics and gene expression experiments are sometimes criticised as being "fishing expeditions," rather than hypothesis-driven. This is not a problem if you are searching for fish, such as new genes which are involved in a given metabolic pathway, or which may be used in a diagnostic or predictive fashion, or as drug targets, as the following selected examples demonstrate. These large-scale experiments can be said, instead, to be hypothesis-generating, after which the researchers must return to single-gene methods for further investigation and functional characterisation of the candidates.

1.7.2.1 Cancer biology and pathogenesis

Large-scale profiling technologies have been used successfully for the identification of target genes which are involved in a given metabolic pathway or biological process. For example, the tumor suppressor gene adenomatous polyposis coli (APC) is inactivated in most colorectal cancers and in gastric carcinomas of distal, intestinal type (see previous Section 1.5.4). APC mutations cause the aberrant accumulation of beta-catenin, which then binds T cell factor-4 (Tcf-4), causing increased transcriptional activation of genes downstream. Global gene expression profiles of colorectal cancer cells were generated by SAGE and used to identify two important genes as targets of this signalling pathway: the c-myc proto-oncogene (He *et al.* 1998) and PPARdelta (peroxisomal proliferator activator receptor delta) (He *et al.* 1999). The promoters of both genes were found to contain beta-catenin/Tcf-4-responsive elements. The ability of PPARdelta to bind to its eicosanoid recognition sequences is disrupted by non-steroidal anti-inflammatory drugs (NSAIDs), which are well-known to suppress colorectal (and gastric) carcinogenesis in both animal models and humans. Thus NSAIDs such as aspirin may inhibit tumorigenesis through inhibition of PPARdelta, the gene for which is normally regulated by APC.

1.7.2.2 Cancer class prediction and discovery

Large-scale profiling technologies have also been used successfully for cancer classification. In some cases, the results simply correlate with and reinforce existing diagnostic categories; this has been termed class prediction. In others, the results go beyond existing knowledge and provide new criteria for improved cancer diagnosis, prognostication and therapy; this has been termed class discovery.

In terms of class prediction, artificial neural networks (ANNs) have been trained to classify the gene expression profiles of the so-called small round blue-cell tumours of childhood, which often cause diagnostic difficulties in clinical practice (Khan *et al.* 2001). Four distinct diagnostic categories were used (which is an over-simplification!). Large-scale gene expression profiles contain a huge amount of data and ANNs represent one way of approaching this bioinformatics problem. After initial training, the ANN correctly classified all samples, including additional blinded samples, and identified the genes most relevant to the classification, which represent new candidate targets for therapy.

Large-scale profiles of gene expression may also reveal similarities and differences between tumours which are not evident by classical morphology. In some cases the results shed light on the underlying tumour biology. For example, small cell lung carcinoma is common, malignant, and generally regarded as being of neuroendocrine origin, as are carcinoid tumours of the lung, which tend to behave in a benign fashion. The two tumour types and normal bronchial epithelial cells were studied using cDNA arrays (Anbazhagan *et al.* 1999). The expression profiles of small cell carcinoma and normal bronchial epithelium were similar but differed from those of the carcinoid tumours, which instead resembled astrocytic brain tumours (high-grade astrocytoma). Small cell lung carcinomas are thus related to, and may originate from, bronchial epithelial cells, whereas carcinoids have more in common with neural crest-derived brain tumors, which would tend to contradict current opinion.

A number of extremely exciting recent studies have investigated groups of tumours which by traditional criteria are homogeneous but which have a variable clinical outome. They have been able to identify previously undetected and undetectable molecular heterogeneity and to characterise gene clusters of prognostic importance. Diffuse large B-cell lymphoma (DLBCL) is the most common subtype of non-Hodgkin's lymphoma but is clinically disparate: 40% of patients respond well to current therapy and survive, whereas the remainder die of their disease. Using DNA microarrays, two distinct forms of DLBCL were identified, with gene expression patterns indicative of different stages of B-cell differentiation (Alizadeh *et al.* 2000). One type expressed genes characteristic of germinal centre B cells; the second type expressed genes normally induced during *in vitro* activation of peripheral blood B cells. The differentially expressed genes reflected variations in the proliferation rate and differentiation state of the tumours, and in the host response. Patients with activated B-like DLBCL had a significantly worse overall survival and presumably should be treated more aggressively *ab initio*. Interestingly, some of the genes characteristic of germinal center B-cells had already been implicated in lymphoma pathogenesis, including bcl-6 and CD10. Follow-up studies have shown that when these genes are studied by immunohistochemistry as individual markers (as with oestrogen receptor in breast cancer), rather than as part of a large profile, they continue to provide prognostic information, enabling the separation of patients into sub-groups with good and bad outcomes and therefore guiding management (Barrans *et al.* 2002).

A similar study focussed on young patients with breast cancer but without lymph node spread at presentation. These women have a poorer prognosis than older patients of similar stage, because they more commonly develop distant metastases, but this event cannot be predicted using traditional clinico-pathological criteria. The risk of metastases is reduced by one-third by chemotherapy, but up to 70% of patients would have survived anyway without it. DNA microarray analysis of the primary tumours of patients who did or did not develop metastatic disease identified an expression signature of 70 genes which predicted subsequent distant spread (van 't Veer *et al.* 2002). This time, well-established individual markers such as oestrogen receptor and c-erbB2 were absent from the list, which instead included genes involved in: cell cycle control; invasion e.g. metalloproteinases; angiogenesis; and signal transduction. Clinical use of such a gene expression profile would enable the direction of chemotherapy to those likely to benefit, avoiding unnecessary toxicity in the remaining patients and saving money.

Such stratification of patients with early stage disease which is homogeneous by traditional staging methods but which has a variable outcome is also possible using DNA technology. For example, imbalances of chromosome 8p and 18q in early-stage colorectal cancers, lacking lymph node or distant metastases, were sought by applying digital SNP to DNA prepared from 180 paraffin-embedded tumours (Zhou *et al.* 2002). The 5-year survival was 100% for patients with tumours without allelic imbalance, but fell to 58% for those with allelic imbalance on both chromosomes. This predictive effect was independent of other variables, such that Duke's stage A tumours (up to tumour stage T2) with allelic imbalance were much more likely to recur than the higher stage Duke's stage B tumours (T3 and over) without allelic imbalance. In colorectal cancer patients without metastasis, therefore, allelic imbalance appears to be a better predictor of prognosis than histopathological stage. To my knowledge, this is the first time that molecular markers have surpassed traditional clinico-pathological staging in a solid tumour.

The study of proteins can also make a significant clinical contribution. Like gastric carcinoma, ovarian cancer presents at a late clinical stage in more than 80% of patients, in whom the five-year survival averages 35% (which is obviously better than gastric carcinoma's 10%). In contrast, the five-year survival of patients with early (stage I) ovarian cancer exceeds 90% and most patients can be cured by surgery alone. Simply increasing the proportion of women diagnosed with early disease would therefore significantly reduce mortality. Such a screening strategy would be essentially the same as that already used for breast cancer and cervical pre-cancer, but the latter are easier to view and sample since the tissues are more accessible. Proteomic technology using mass spectroscopy spectra has been able to identify a proteomic pattern that discriminates patients with ovarian cancer, including stage I disease, from those without (Petricoin *et al.* 2002), with 100% sensitivity and 95% specificity. Proteomics thus offer the extremely exciting prospect of a screening tool for ovarian cancer. Other carcinomas, including gastric and pancreatic, should be equally amenable to a proteomic approach for screening/early diagnosis.

Molecular classification of tumours on the basis of large-scale genetic profiling, at the DNA, RNA and protein levels, can thus identify previously undetected and clinically significant subtypes of cancer. This molecular classification is likely to supplement traditional microscopic diagnosis and to alter clinical practice.

1.8 Large-scale expression profiling of normal stomach and gastric carcinoma by SAGE

By this stage, I hope that you are convinced that the stomach and gastric carcinoma are worthwhile subjects for study. Gastric carcinoma is a major cause of cancer death world-wide. Recent diagnostic and therapeutic advances have been few but these should be expedited by improved understanding of its biology, as has been the case with, for example, breast and colorectal cancer.

The classical molecular studies of gastric carcinoma have been extremely fruitful, for example in the identification of the importance of E-cadherin in diffuse cancers, but have of necessity focussed on abnormalities in single or small numbers of genes. Changes in the global pattern of gene transcription may be equally or more significant functionally and clinically, and technological advances now permit their investigation. Since this project was started, gastric carcinoma has been studied by differential display (DD) (Ebert *et al.* 2000; Jung *et al.* 2000; Yoshikawa *et al.* 2000; Wang *et al.* 2001) and Clontech cDNA membrane arrays (El-Rifai *et al.* 2001). These studies are discussed alongside the results of this research in Chapter 4.

However, my chosen method was serial analysis of gene expression (SAGE), which produces comprehensive, quantitative and reproducible gene expression profiles. When I started this project, only two papers had been published on SAGE, both from the Johns Hopkins group: the original description (Velculescu *et al.* 1995); and its first major application, which was profiling and comparing genes expressed in colorectal and pancreatic carcinomas and their respective normal tissues (Zhang, L. *et al.* 1997). Since then, the applications of SAGE have been many and varied. SAGE has provided an encyclopaedic description of mRNA transcripts in yeast (Velculescu *et al.* 1997), enabling direct comparison with the yeast genome. It has also been used to investigate carcinogenesis at more basic levels, such as p53 function and p53-induced apoptosis (Madden *et al.* 1997; Polyak *et al.* 1997), as well as targets of APC (He *et al.* 1998; He *et al.* 1999) (see Section 1.7.2.1). Within the field of clinical cancer research, and since the start of this project, SAGE has been applied to other common adenocarcinomas as well as colon and pancreas: breast (Nacht *et al.* 1999), ovary (Hough *et al.* 2000), prostate (Waghray *et al.* 2001), lung (Nacht *et al.* 2001), and pancreas again (Argani *et al.* 2001; Ryu *et al.* 2001), as well as to squamous lung carcinoma (Hibi *et al.* 1998) and other cancers (Lal *et al.* 1999; Nacht *et al.* 1999; Velculescu *et al.* 1999). However, to my knowledge, no SAGE data have yet been published, outwith the web, for carcinoma of the stomach. Because SAGE is labour-intensive and hence limited to small numbers of specimens, resulting candidate genes are usually validated in a larger set of samples (Velculescu *et al.* 1995; Zhang, L. *et al.* 1997; Velculescu *et al.* 2000).

1.8.1 Aim

My aim was therefore to identify and compare genes expressed in gastric adenocarcinoma and in normal stomach using SAGE

1.8.2 Objectives

My specific objectives were:

1. **To establish the method of SAGE in this laboratory.**
2. **To examine gastric carcinoma and normal stomach using SAGE.** Since the most common subtype of gastric carcinoma is distal and intestinal (Fuchs *et al.* 1995), two such tumours were studied using SAGE, along with one sample of normal corresponding normal distal, antral, mucosa.
3. **To obtain and analyse the gastric gene expression profiles.** Around 10,000 tags were obtained for each sample. Catalogues of the genes expressed were assembled.
4. **To compare the gastric carcinomas with each other, with normal stomach and, on-line, with other adenocarcinomas and glandular mucosae.** Transcripts were identified which were: abundant; differentially expressed between normal and tumour stomach; or gastric-specific by comparison with the on-line libraries.
5. **To validate the SAGE profiles by further study of candidate genes.** Selected genes were then validated in a panel of 19 gastro-intestinal normal and tumour tissues and cell lines by Northern blotting and immunohistochemistry.
6. **To study in more detail genes identified as showing differential expression, particularly where these are novel or little characterised sequences.** One gene was new and highly expressed in the stomach, and I went on to characterise it in detail at the mRNA and DNA levels in human and model organisms.

Figure 1.1: Gross anatomy of stomach

This figure shows the normal stomach and oesophagus. The oesophagus is the tube-like structure at the top. It is lined by stratified squamous epithelium and so appears white. In contrast, the stomach is lined by simple columnar epithelium organised in glands within a mucosa and is pale pink/brown in colour. Although by gross examination normal gastric mucosa appears rather uniform, microscopically it can be divided into three types, which correspond to different locations in the stomach. The cardia comprises the first 1-2 cm of the stomach beyond the oesophago-gastric junction. The bulk of the stomach then comprises body mucosa which secretes acid and proteases (see Section 1.2.2). The distal stomach is made up of the antrum and pylorus, which have a similar mucosa. Thereafter the stomach joins the duodenum. The concave border of the stomach on the left is termed the lesser curvature, and this is the site of most gastric carcinomas (see Sections 1.4.4 and 1.4.7). The longer, convex, border of the stomach on the right is the greater curvature.

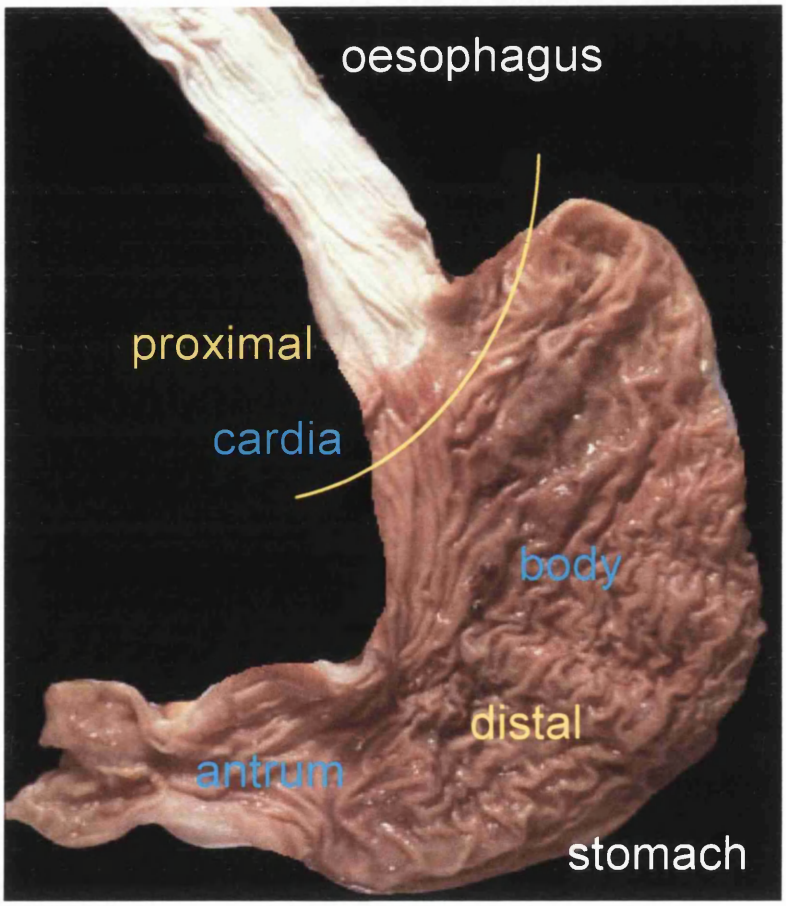
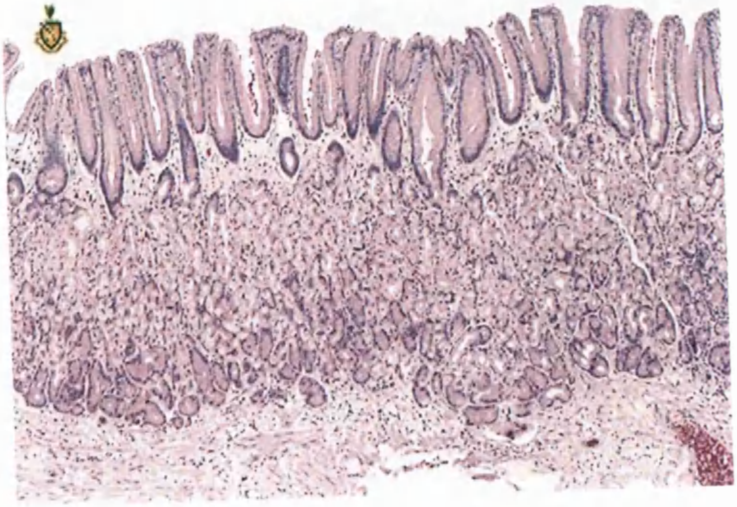

oesophagus
proximal
cardia
body
distal
antrum
stomach

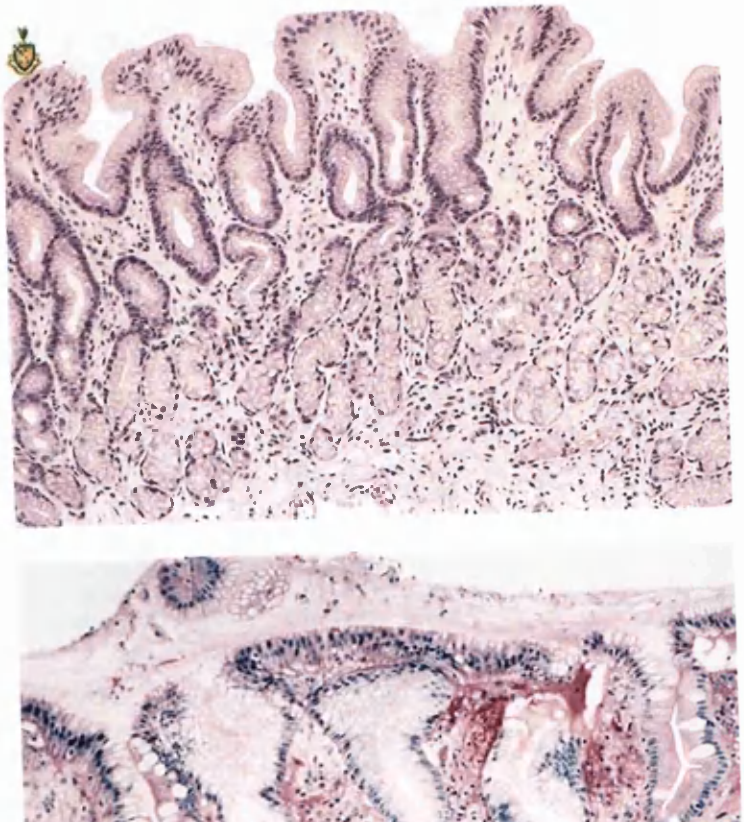
Figure 1.2: Histology of normal gastric antral and body mucosa and of intestinal metaplasia and atrophy

The stomach is divided anatomically and by its mucosa into three distinct areas, all lined by simple glandular epithelium (see Section 1.2.2). The picture at the top shows normal gastric body mucosa. This lines most of the stomach and comprises long, tubular glands lined by a variety of epithelial cells. Within each gland, the upper quarter is lined by mucous (mucus-secreting) cells of surface / (glandular) neck type. The next cells form the stem cell (proliferative) compartment. The next quarter of the gland crypt contains acid-producing parietal cells, which secrete hydrogen and chloride ions into the gastric lumen. Parietal cells have a very high energy requirement and therefore contain large numbers of mitochondria, which yield a pink granularity to their microscopic appearance on routine histological staining with haematoxylin and eosin (H&E). In the lowermost quarter of the gland are the gastric chief, or peptic, cells which secrete the main gastric enzyme, pepsin, in its precursor form, pepsinogen. Since their main role is to produce a protein for secretion in abundance, these cells contain large quantities of rough endoplasmic reticulum and hence stain dark purple with H&E staining, as seen in the picture.

The picture in the middle shows normal gastric antral mucosa, which lines the distal part of the stomach. Gastric antral mucosa is simpler than that in the body, being composed mainly of mucus-secreting glands. There are also endocrine cells: the most numerous in the antrum secrete gastrin and hence are termed G-cells, but these are not clearly seen with routine H&E staining, instead being better demonstrated by immunohistochemistry (see Section 4.1.2.2).

The picture at the bottom shows gastric mucosa with intestinal metaplasia and atrophy. The mucosa is rather thinner than normal (atrophy literally means lack of growth). There is also a mild chronic gastritis, composed of a scattered infiltrate of lymphocytes, which are the small cells staining dark blue clustered around one of the glands in the middle of the picture. There are two areas of complete (Type I) intestinal metaplasia (IM) (see Section 1.3.1), one each on the left and right of the picture. The superficial and foveolar (pit) regions of the gastric glands are normally covered with mucous-secreting columnar cells with clear cytoplasm. In IM, this is replaced by pink-staining absorptive columnar cells with a microvillous brush border and interspersed goblet cells, characteristic of normal small intestinal epithelium. IM and atrophy predispose to the development of distal gastric carcinoma of intestinal histological type (see Section 1.4.6).





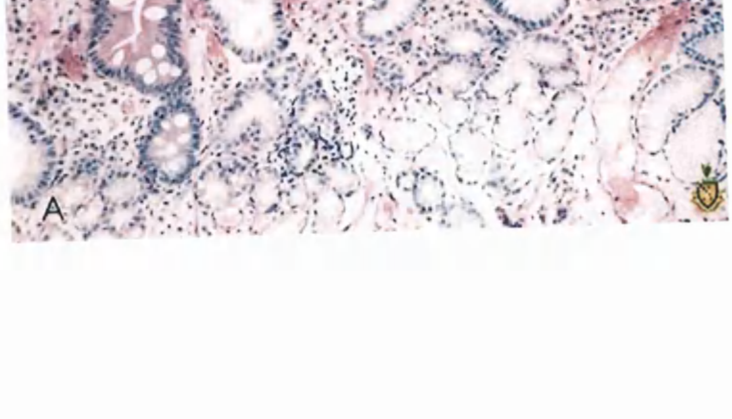
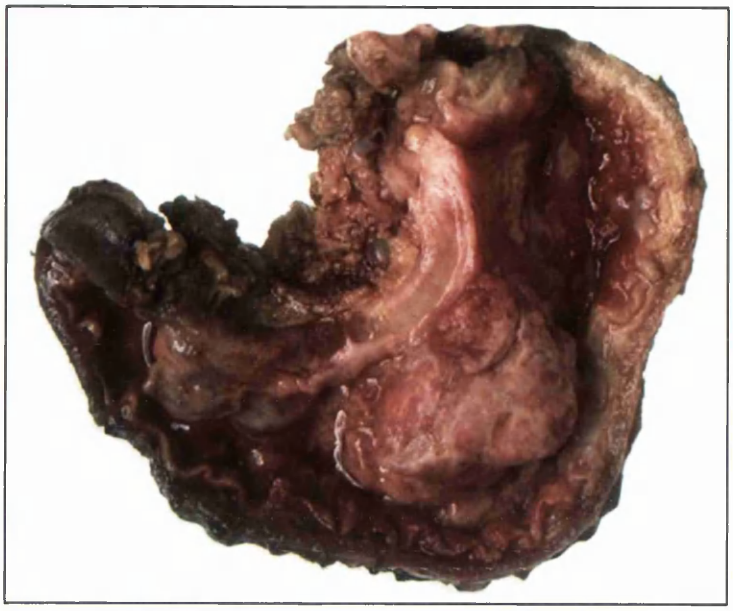

A

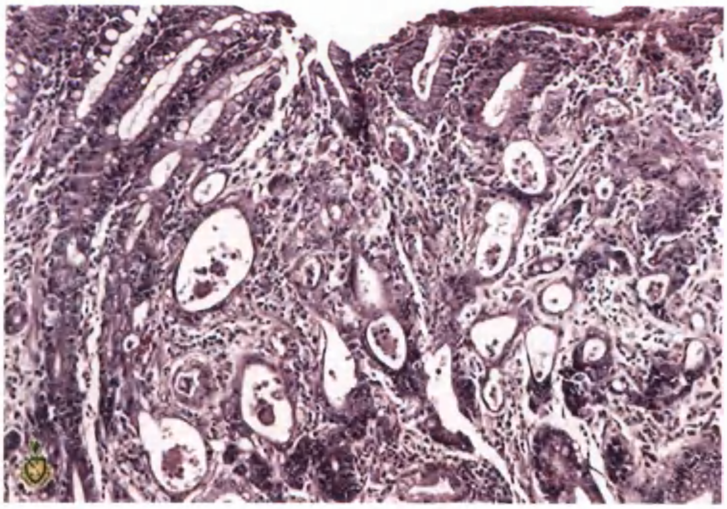
Figure 1.3: Gross pathology of gastric carcinoma and histology of the intestinal and diffuse subtypes

The top picture shows a gastrectomy specimen from a surgical resction for gastric carcinoma. The tumour lies in the mid stomach (predominantly body) and involves much of its circumference. As is usual, the cancer has grown to a large size before coming to clinical attention. Staging of the tumour involves assessing: the extent of its spread locally through the wall of the stomach; and tumour invasion of lymph nodes, mainly in the fatty lesser omentum along the lesser curve towards the upper left of the specimen, and in the greater omentum, on the opposite side of the stomach (see Sections 1.4.4 and 1.4.7). Obviously, staging also involves the identification or exclusion of metastatic spread.

The middle picture shows a gastric carcinoma of intestinal type, which is the focus of this project (see Section 1.4.5). The normal mucosa is replaced by infiltrating tumour in the form of cohesive glandular structures, which recapitulate the normal mucosal architecture: these tumours thus tend to be well or moderately differentiated. In this case, within the tumour tubules, pink material is seen: this is apoptotic debris. Both the glandular appearance and presence of prominent cell death is reminiscent of colonic carcinoma.

The bottom picture shows a gastric carcinoma of diffuse type (see Section 1.4.5). The mucosa is diffusely infiltrated by individual, non-cohesive malignant cells which are often described as having a signet-ring appearance: the cells contain a mucin vacuole which is clear on routine H&E staining and which pushes the nucleus to one side. The malignant cells bear almost no resemblance to normal stomach and therefore these tumours are usually regarded as poorly differentiated *ab initio*. The superficial and foveolar regions of the gastric mucosa appear relatively normal, without gastritis or intestinal metaplasia.





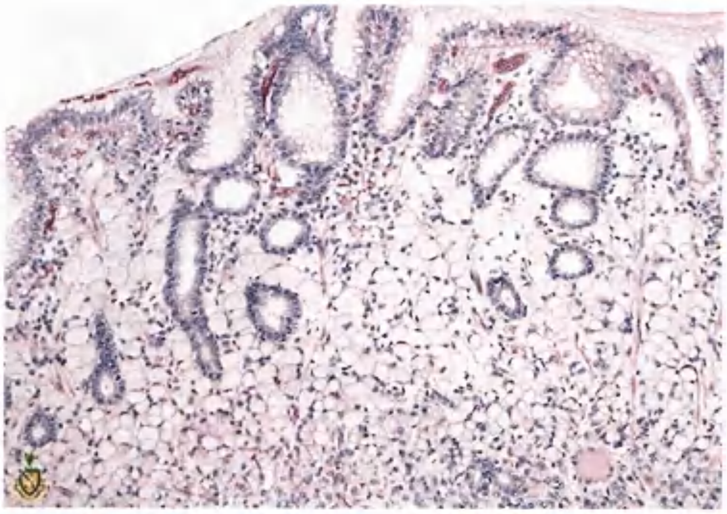


Table 1.1: Current TNM staging system for gastric carcinoma

This table describes the current TNM (Tumour, Node and Metastasis) staging system for gastric adenocarcinoma (Klein Kranenbarg *et al.* 2001). The nodal stage is currently defined according to the number of lymph nodes involved: previously the stage related also to their distance from the primary tumour.

TNM classification of staging of gastric carcinoma

Primary tumour

TX	primary tumour cannot be assessed
Tis	carcinoma *in situ*
T1	tumour invades lamina propria or submucosa
T2	tumour invades muscularis propria (muscle coat) or subserosa
T3	tumour penetrates the serosa (i.e. involves the lining of the peritoneal cavity); no invasion of adjacent structures
T4	tumour invades adjacent structures (e.g. pancreas, colon)

Regional lymph node metastases

NX	nodes cannot be assessed
N0	no nodal metastases
N1	1-6 involved nodes
N2	7-15 involved nodes
N3	16 or more involved nodes

Distant metastases

MX	distant metastases cannot be assessed
M0	no distant metastases
M1	distant metastases

Overall tumour stage

0	Tis	N0	M0
I	T1	N0-1	M0
	T2	N0	M0
II	T1	N2	M0
	T2	N1	M0
	T3	N0	M0
III	T2	N2	M0
	T3	N1-2	M0
	T4	N0-1	M0
IV	T4	N2	M0
	Any T	Any N	M1

Table 1.2: Survival after resection of gastric carcinoma in US and Japanese patients

This table describes patients who underwent resection for gastric carcinoma in the US and Japan (Fuchs *et al.* 1995). The relative proportions at each tumour stage and the 5-year survival is compared.

The US cases are based on data on 11,087 patients from 700 US hospitals. The age-adjusted survival is based on the 10,237 patients who underwent gastric resection. The Japanese cases are based on data on 3,176 patients who underwent gastric resection at the National Cancer Centre Hospital, Tokyo, Japan.

Stage I tumours are more common in Japan. This may relate partly to differences in diagnostic criteria in pathological evaluation: lesions that in the West would be called high-grade dysplasia would be called carcinoma *in situ* in Japan and hence would increase the relative proportion of early tumours (Schlemper *et al.* 1997). In addition, endoscopic screening programs in Japan make the pick-up and treatment of early tumours more likely.

However, the relative proportions of tumours of Stage II, III and IV are similar in both populations: the ratio is essentially 1:2:2. Stage-for-stage, patients in the West have a much poorer prognosis in terms of comparable 5-year survivals. Possible explanations include: surgery which is technically less good and less extensive in the West; biological differences in disease and/or patient population; and relative under-staging in the West, such that patients who are thought to be, say, Stage II, are actually Stage III (for example, with a T2 primary tumour and lymph node metastases which were present, but not identified because of inadequate dissection).

It is worth remembering that, in the West, most patients with gastric carcinoma do not undergo resection because their tumours at presentation are too far advanced, thus they would not be included in this surgical datatset.

Tumour	**United States (1982-1987)**		**Japan (1971-1985)**	
stage	**No. of cases (%)**	**5-year survival (%)**	**No. of cases (%)**	**5-year survival (%)**
I	2004 (18.1%)	50%	1453 (45.7%)	90.7%
II	1796 (16.2%)	29%	377 (11.9%)	71.7%
III	3945 (35.6%)	13%	693 (21.8%)	44.3%
IV	3342 (30.1%)	3%	653 (20.6%)	9%

Table 1.3: Molecular genetic changes in gastric carcinoma

This table summarises the molecular events so far identified in gastric adenocarcinoma (reviewed in (Tahara *et al.* 1996; Younes *et al.* 1997)). The genes involved are listed and sub-divided by type. The carcinogenic change is described. For the tumours, the data are separated by the two histological subtypes of gastric adenocarcinoma: intestinal and diffuse. The figures are percentages, and where results have varied, a range is given.

The premalignant conditions described relate to tumours of intestinal type only. Intestinal metaplasia (IM) has been described previously in Section 1.3.1: it is a consequence of chronic gastritis and is often accompanied by gastric atrophy. Adenomas represent a stage further down the carcinogenic pathway (Correa 1992); by definition, their epithelium shows dysplasia. Clearly defined adenomas are very rare in areas of low incidence of gastric adenocarcinoma, such as the West, but are more common in the Far East and South America. However, even in the West, areas of dysplastic epithelium are common in resection specimens adjacent to the main tumours.

The abbreviated names are: RER+, replication error positive; MI, microsatellite instability; APC, adenomatous polyposis coli; MCC, mutated in colon cancer; DCC, deleted in colon cancer; EGF, epidermal growth factor; EGFR, epidermal growth factor receptor; and TGFalpha, transforming growth factor alpha. LOH is loss of heterozygosity and IM is intestinal metaplasia.

It is obvious that some alterations are common to both tumour types, and sometimes to the pre-malignant conditions, for example, microsatellite instability, loss of normally functioning p53 and activation of telomerase. Other changes, such as APC and K-ras mutations, are restricted to tumours of intestinal type and their precursors. In contrast, E-cadherin and K-sam mutations are limited to diffuse carcinomas. This provides further evidence for different carcinogenic mechanisms for the two tumour types, as well as supporting the pathway proposed by Correa leading to intestinal tumours via IM (and dyplasia).

Gene	Alteration	Intestinal tumours	Diffuse tumours	Premalignant conditions (for intestinal tumours)
Microsatellite instability				
RER+, MI	acquisition of phenotype	17-40%	18-64%	42% adenomas, 33% IM
Tumour suppressor genes				
p53	LOH, mutation	60-70%	75-76%	22-30% adenomas, 14-40% IM
APC	LOH, mutation	40-80%	0%	42% adenomas, 6% IM
MCC	LOH	up to 100%	0%	
DCC	LOH	40-60%	0%	
E-cadherin	LOH, mutation	0%	50%	
Oncogenes				
K-ras	point mutation	9-18%	0%	yes, including IM
c-erbB2	amplification	20%	0%	
EGF	over-expression	25-40%	20%	
EGFR	over-expression	15-50%	25%	
TGFα	over-expression	60%	55%	
K-sam	amplification	0%	33%	
c-met	amplification	19%	39%	yes
Cyclins, cyclin-dependent kinases (CDKs), their inhibitors and related molecules				
cyclin E	over-expression	57%	63%	
cyclin E	amplification	33%	7%	
p21	over-expression	77%	76%	
p16	loss	12%	31%	
$p27^{KIP1}$	loss	46-60%	69%	10% adenomas
bcl-2	over-expression	88%	7%	
Metastasis-associated gene				
CD44	aberrant transcript	100%	100%	
Telomerase				
telomerase	activation	100%	90%	50% adenomas, over 23% IM

2 Materials and Methods

2.1 Materials

2.1.1 Equipment

Equipment which is standard in most laboratories and which was used in this work is not separately listed by supplier below, but included: wet and dry ice; water baths; 37°C incubators; vortex mixer; refrigerators; -20°C and -70°C freezers; microwave oven; sterile and non-sterile glass pipettes, bottles, flasks and beakers; plastic bottles, beakers, measuring cylinders and boxes; aluminium foil, cling film and plastic wrapping; platform shaker; magnetic stirrer and hotplate; metalware including forceps and spatulas; and facilities for culture of mammalian cells, for microbiological culture, for working with radioactive isotopes and for the development of X-ray film.

Anachem Ltd (Luton, UK)

Gilson PIPETMAN® pipettes (up to 10, 20, 200 and 1000 µl, with different sets for RNase-free, pre-PCR and general use)

Applied Biosystems Ltd (Warrington, UK)

ABI Prism Automated Sequencer (Models 373 and 377)

B Braun (Melsungen, Germany)

Mikrodismembrator II plus and bead mill (for powdering of frozen tissue)

Beckman (RIIC) Ltd (High Wycombe, UK)

Centrifuge tubes

Microfuge® R centrifuge (refrigerated microcentrifuge)

GS-6R centrifuge (low-speed refrigerated centrifuge)

Avanti™ J-25 centrifuge (high-speed refrigerated centrifuge)

BioRad Laboratories Ltd (Hemel Hempstead, UK)

Power supply (pack)

Sub-Cell® GT apparatus for agarose gel electrophoresis (for Northern blotting)

Gel Doc 1000 gel documentation system

Charles Austin Pumps Ltd (Byfleet, UK)

Capex L2C vacuum pump (for use with QIAvac 6S)

Genetic Research Instrumentation (Braintree, UK)

Atto Maxi Slab for vertical PAGE: gel tank and pouring apparatus with 160 x 160 mm glass gel plates, 1.5 mm spacers and 12 and 20 well combs

Millipore (Watford, UK)

Milli-Q plus PF water purification system

MJ Research Inc (Watertown, MA, USA)

PTC-100^{TM} Programmable Thermal Controller (for sequencing PCRs)

Pharmacia Biotech Ltd (St Albans, UK)

GeneQuant RNA/DNA Calculator (spectrophotometer)

Qiagen Ltd (Crawley, UK)

QIAvac 6S (for large-scale DNA purification with QIAquick 8 PCR Purification Kit)

Stratagene Ltd (Cambridge, UK)

UV Stratalinker 2400 (for UV cross-linking)

Thermo Hybaid (Ashford, UK)

Cell Shock Electroporator and electroporation cuvettes

Electro-4 system (for standard and large-scale agarose gel electrophoresis)

Glass roller bottles (for membrane hybridisation)

Hybridisation oven with rotisserie

PCR Thermal Cyclers: Touchdown and Omnigene machines with hot-lids

Omnislide Thermal Cycler (for *in situ* hybridisation)

Treff Lab (Degersheim, Switzerland)

Adjustable multichannel pipettes (for large-scale DNA purification and gel electrophoresis)

X-ray Accessories Ltd (Bushey, UK)

X-ray film cassettes

2.1.2 General plasticware

ABgene (Epsom, UK)

OmniTube PCR plates with 96 x 0.3 ml tubes, plus caps

Becton Dickinson UK Ltd (Oxford, UK)

Falcon conical tubes (15 and 50 ml)

Falcon snap-cap round-bottomed tubes (14 ml)

Needles (21 gauge, “green”)

Plastic pipettes (5, 10 and 25 ml)

Syringes (between 2 and 10 ml)

Bibby-Sterilin Ltd (Stone, UK)

Bijou tubes (5 ml)

Petri dishes (90 and 140 mm diameter)

Universal tubes (25 ml)

Corning Costar UK (High Wycombe, UK)

Costar® cell scrapers

Spin-X filter microcentrifuge tubes

Elkay Laboratory Products (UK) Ltd (Basingstoke, UK)

Microcentrifuge tubes (0.2, 0.5, 1.5 and 2.0 ml) (both flip-cap and screw-cap)

Standard pipette tips

Greiner Labortechnik Ltd (Stonehouse, UK)

Aerosol-resistant pipette tips (10, 20 200 and 1000 μl) (for RNA and pre-PCR work)

Invitrogen Life Technologies Ltd (Paisley, UK)

Nunclon™ tissue culture flasks (12.5, 25 and 75 ml)

Swann-Morton (Sheffield, UK)

Sterile disposable scalpels

2.1.3 Chemicals

Stock solutions were made up with dH_2O. For RNA work, the dH_2O used was treated with diethylpyrocarbonate (DEPC) (Section 2.2.2.1).

Braun Medical Ltd (Sheffield, UK)

Water for Injections BP

Fisher Scientific UK Ltd (Loughborough, UK)

Acetic acid, glacial

Ammonium acetate (10 M stock solution)

Ammonium persulfate (APS) (10% stock solution)

Ammonium sulphate ($(NH_4)_2SO_4$) (1 M stock solution)

Chloroform

Diaminoethanetetra-acetic acid disodium salt (EDTA) (0.5 M stock solutions at pH 7.5 and pH 8.0)

Glycerol

Hydrochloric acid

Isopropanol (isopropyl alcohol)

Magnesium chloride ($MgCl_2$) (1 M stock solution)

Methanol

Propan-2-ol (isopropanol)

Sodium chloride (5 M stock solution)

Sodium dodecyl sulphate (SDS) (10% stock solution)

Sodium hydroxide (0.5 M stock solution)

Trichloroacetic acid (100% w/v stock solution)

Tris HCl

Tris base

Xylene

Hayman Ltd (James Borrough) (Witham, UK)

Absolute alcohol (ethanol) (analytical reagent grade) (100% and 70% stock solutions)

Invitrogen Life Technologies Ltd (Paisley, UK)

Tris base (1 M stock solutions at pH 7.5 and 8.8)

Sigma (Dorset, UK)

Ampicillin

β-mercaptoethanol

Bromophenol Blue

Calcium chloride

Copper sulphate

3,3-Diaminobenzidine (DAB)

Diethylpyrocarbonate (DEPC)

N,N -Dimethylformamide (DMF)

Ethidium bromide (10 mg/ml stock solution)

Hydrogen peroxide

Mineral oil

N,N,N',N'-Tetramethylethylenediamine (TEMED)

Xylene Cyanol

2.1.4 Buffers and other solutions

Buffers and other solutions were made up with dH_2O. For RNA work, the dH_2O used was treated with diethylpyrocarbonate (DEPC) (Section 2.2.2.1).

LoTE

3 mM Tris-HCl (pH 7.5), 0.2 mM EDTA (pH 7.5)

TE

10 mM Tris-HCl (pH 7.5), 1 mM EDTA (pH 7.5)

2X Binding and Washing (B+W) buffer (for use with Dynal magnetic beads)

2 M NaCl, 10 mM Tris-HCl (pH 7.5), 1 mM EDTA (pH 7.5)

1X B+W buffer

1 M NaCl, 5 mM Tris-HCl (pH 7.5), 0.5 mM EDTA (pH 7.5)

50X Tris-Acetate-EDTA (TAE) buffer

242 g Tris base, 57.1 ml glacial acetic acid and 100 ml 0.5 M EDTA (pH 8.0) made up to 1 litre with dH_2O

Buffers in a Box, Invitrogen Life Technologies Ltd (Paisley, UK)

10X Tris-Borate-EDTA (TBE) buffer

1 M Tris, 0.9 M Boric Acid, 0.01 M EDTA

20X Salt and Sodium Citrate (SSC) buffer

3 M NaCl, 0.3 M Sodium Citrate

Phosphate Buffered Saline (PBS)

2.1.5 Tissue, cell and RNA samples

2.1.5.1 Primary tissues for SAGE, Northern blotting and immunohistochemistry

The primary tumours and normal mucosa used for SAGE and its validation by Northern blotting and immunohistochemistry (Chapter 4) were collected from oesophagus, stomach and colon, as described in Section 2.2.1. The gastric adenocarcinomas are detailed in Table 2.1. The other specimens comprised: normal gastric mucosa from the cardia, body and antrum; normal squamous oesophageal mucosa; oesophageal squamous carcinoma; normal colonic mucosa; and colonic adenocarcinoma.

2.1.5.2 Primary tissues for *in situ* hybridisation

The tissues described above and the cultured cells described below were used for *in situ* hybridisation for the new gastric gene identified via SAGE. Two other sets of tissues were also tested, as described in Section 2.2.1 and discussed in Chapter 5.

The first was a range of further normal and diseased upper gastro-intestinal tissues. Gastric samples included: normal mucosa; gastritis; intestinal metaplasia and atrophy; gastric ulceration; and gastric carcinoma. Oesophageal samples included: normal mucosa; Barrett's oesophagus; and oesophageal squamous carcinoma and adenocarcinoma. Duodenal samples included: normal mucosa; gastric metaplasia; and duodenal ulceration. For each condition, between two and five different blocks were examined.

The second set was a wide range of gynaecological, glandular, solid epithelial and other tissues, both normal and tumorous. Gynaecological samples included: normal cervix; cervical squamous carcinoma and adenocarcinoma; normal uterus including endometrium; endometrial adenocarcinoma; fallopian tube; normal ovary; ovarian serous and mucinous cystadenoma and adenocarcinoma; placenta and peritoneum. Glandular and solid epithelial tissues included: salivary gland; normal tissue and adenocarcinoma from bile duct, pancreas, colon, breast and prostate; appendix; normal lung; squamous carcinoma, adenocarcinoma and small cell carcinoma of lung; normal thyroid and follicular carcinoma; kidney; normal testis and teratoma; and skin including adnexal glands. Other tissues included: heart; thymus; lymph node; tonsil; brain; cerebellum; malignant melanoma; skeletal muscle and bone. For the gynaecological and glandular tissues, between two and five different tissue blocks were examined; for the others, one or two blocks were used.

2.1.5.3 Cultured cells

For the cultured cells listed below, cells described as adherent are derived from adenocarcinomas of intestinal type whereas those described as spherical originate from diffuse tumours.

Deutsche Sammlung von Mikroorganismen und Zellkulturen (Braunschweig, Germany)

MKN-45

Derived from a poorly differentiated gastric adenocarcinoma in a female aged 62. Mainly spherical cells.

European Collection of Cell Cultures (Wiltshire, UK)

AGS

Derived from a gastric adenocarcinoma in a Caucasian female aged 54. Adherent cells.

KATO-III

Derived from a metastasis of a gastric carcinoma in a male aged 55. Spherical cells.

OE19

Derived from a moderately differentiated adenocarcinoma of gastric cardia/oesophago-gastric junction (proximal) in a male aged 72. Adherent cells.

Invitrogen Life Technologies Ltd (Paisley, UK)

L-Glutamine
Nutrient Mixture F-12 (Ham)
Penicillin/Streptomycin Solution
RPMI 1640 Medium
Trypsin

Autogen Bioclear UK Ltd (Calne, UK)

Foetal Bovine Serum (FBS)

2.1.5.4 Purchased RNA

BD Clontech UK (Basingstoke, UK)

Human Stomach Poly(A) RNA (5 μg pooled from whole, normal, stomach of four Caucasian males and females with sudden deaths)

2.1.5.5 Purchased Northern blots

Invitrogen Life Technologies Ltd (Paisley, UK)

Northern Territory™ Human Normal Tissue Blot III, bearing total RNA from Tonsil, Thymus, Appendix, Lymph Node, Gallbladder, Prostate, Testis and Ovary, loaded at 20 μg per lane

OriGene Technologies, Inc (Rockville, MD, USA)

Multiple Choice™ Northern Blot, bearing total RNA from Spleen, Liver, Colon, Stomach, Testis and Placenta, loaded at 20 μg per lane

2.1.6 RNA and DNA purification, cDNA synthesis and magnetic beads

Ambion (Houston, TX, USA)

DNAZap® (PCR DNA Degradation Solution)

Poly(A)Pure™ Kit

Phenol/chloroform/isoamyl alcohol (25:24:1, v/v/v) (hereafter called phenol/chloroform (P/C))

RNaseZap® (RNase Decontamination Solution)

Dynal UK Ltd (Bromborough, UK)

Dynabeads® mRNA DIRECT kit containing Dynabeads® Oligo(dT)$_{25}$ (oligo(dT)-coated magnetic beads)

Dynabeads® M-280 Streptavidin (streptavidin-coated magnetic beads)

Dynal sample mixer

Magnetic Particle Concentrator (magnetic stand to immobilise beads)

Invitrogen Life Technologies Ltd (Paisley, UK)

cDNA Synthesis System (now superceded by the SUPERSCRIPT™ Choice System for cDNA Synthesis)

TRIZOL® Reagent

Miltenyi Biotec Ltd (Bisley, UK)

mRNA Isolation Kit, plus mini MACS (magnetic cell sorting) separator

Promega UK Ltd (Southampton, UK)

Promega MagneSphere® Magnetic Separation Products

Qiagen Ltd (Crawley, UK)

QIAquick® Nucleotide Removal Kit

QIAquick® PCR Purification Kit

QIAquick® 8 PCR Purification Kit

RNeasy® Midi Kit

Roche Diagnostics Ltd (Lewes, UK)

Glycogen, molecular biology grade (20 mg/ml)

2.1.7 Restriction and modifying enzymes

New England Biolabs (UK) Ltd (Hitchin, UK)

Nla III (10 U/µl), supplied with NEBuffer 4 & 100X bovine serum albumin (BSA)

BsmF I (2 U/µl)

Sph I (5 U/µl), supplied with NEBuffer 2 & 100X BSA

T4 Polynucleotide Kinase (10U/µl)

T4 DNA Ligase (4 U/µl and 20 U/µl)

T4 DNA Polymerase (3 U/µl)

DNA Polymerase I Large Fragment (Klenow) (5 U/µl)

Clontech Laboratories (Palo Alto, CA, USA)

Calf Intestinal Alkaline Phosphatase (2.7 U/µl)

Stratagene (CB Amsterdam Zuidoost, The Netherlands)

Pfu DNA Polymerase from Stratagene PCR Polishing Kit

2.1.8 Oligonucleotides

2.1.8.1 Oligonucleotides for SAGE

Oswel Research Products Ltd (Southampton, UK)

The biotinylated oligonucleotides and the linkers were obtained gel-purified from the manufacturers. The working concentration of all SAGE primers was 350 ng/μl.

Biotinylated oligo dT
5' [biotin] T_{18}

SAGE Linker 1A
5' TTT GGA TTT GCT GGT GCA GTA CAA CTA GGC TTA ATA GGG ACA TG 3'

SAGE Linker 1B
5' TCC CTA TTA AGC CTA GTT GTA CTG CAC CAG CAA ATC C [amino mod. C7] 3'

SAGE Linker 2A
5' TTT CTG CTC GAA TTC AAG CTT CTA ACG ATG TAC GGG GAC ATG 3'

SAGE Linker 2B
5' TCC CCG TAC ATC GTT AGA AGC TTG AAT TCG AGC AG [amino mod. C7] 3'

SAGE Primer 1
5' GGA TTT GCT GGT GCA GTA CA 3'

Biotinylated SAGE Primer 1
5' [biotin] GGA TTT GCT GGT GCA GTA CA 3'

SAGE Primer 2
5' CTG CTC GAA TTC AAG CTT CT 3'

Biotinylated SAGE Primer 2
5' [biotin] CTG CTC GAA TTC AAG CTT CT 3'

Invitrogen Life Technologies Ltd (Paisley, UK)

M13 Forward Primer
5' GTA AAA CGA CGG CCA GT 3'

M13 Reverse Primer
5' GGA AAC AGC TAT GAC CAT G 3'

2.1.8.2 Oligonucleotides for SAGE validation and investigation of new gastric gene

Table 2.2 lists the oligonucleotide primers used to validate, by RT-PCR, selected candidate genes which were identified by SAGE. Primers used for the investigation of the new gastric gene, foveolin, are listed in Table 2.3.

2.1.9 Polymerase chain reaction (PCR)

2.1.9.1 Own PCR protocol

Qiagen Ltd (Crawley, UK)
HotStarTaq^{TM} DNA Polymerase Kit
Taq PCR core kit

2.1.9.2 Johns Hopkins' PCR protocol

Invitrogen Life Technologies Ltd (Paisley, UK)
PLATINUM® *Taq* DNA Polymerase (5 U/μl)
10mM dNTP mix (10 mM each dATP, dGTP, dCTP and dTTP at neutral pH)

10X Johns Hopkins PCR buffer
166 mM $(NH_4)_2SO_4$, 670 mM Tris (pH 8.8), 67 mM $MgCl_2$, 100 mM β-mercaptoethanol

Sigma (Dorset, UK)
Dimethyl sulphoxide

2.1.10Agarose and polyacrylamide gel electrophoresis

Invitrogen Life Technologies Ltd (Paisley, UK)
Agarose, electrophoresis grade
10 and 100 bp DNA ladders
0.24-9.5 kb RNA Ladder

Gel loading buffer
30% glycerol, 70% dH_2O and 0.25% (a pinch) of Bromophenol Blue and Xylene Cyanol

Roche Diagnostics Ltd (Lewes, UK)

DNA Molecular Weight Marker V (8-587 bp)

DNA Molecular Weight Marker VIII (19-1114 bp)

Molecular Probes Europe BV (Leiden, The Netherlands)

SYBR® Green I nucleic acid gel stain

Severn Biochem Ltd (Kidderminster, UK)

40% Polyacrylamide (37.5:1 acrylamide:bis)

40% Polyacrylamide (19:1 acrylamide:bis)

2.1.11 Cloning of SAGE concatemers

Invitrogen Life Technologies Ltd (Paisley, UK)

Zero Background™ Cloning Kit

ELECTROMAX™ DH10β™ Cells

S.O.C. Medium

Isopropylthio-β-D-galactoside (IPTG) (2 M stock solution)

X-gal (5-bromo-4-chloro-3-indolyl-b-D-galactoside) (100 mg/ml stock solution in N,N-dimethylformamide)

Low salt LB (Luria-Bertani) agar plates

1% Tryptone, 0.5% Yeast Extract, 0.5% NaCl, 1.5% Agar, pH 7.5, with 50 µg/ml Zeocin™ and 1 mM IPTG, prepared according to instructions in Zero Background™ Cloning Kit

LB agar plates

1% Tryptone, 0.5% Yeast Extract, 1% NaCl, 1.5% Agar, pH 7.0 (prepared by Beatson Labs Central Services) with 50 µg/ml ampicillin, 1 mM IPTG and 40 µg/ml X-gal

Difco Becton Dickinson (Oxford, UK)

Bacto™ Agar

Tryptone peptone

Yeast extract

2.1.12Sequencing

PerkinElmer from Applied Biosystems, Warrington, UK

BigDye Primer Kit

2.1.13Bioinformatics

2.1.13.1 Web-sites

SAGE web-site at Johns Hopkins University (Baltimore, MD, USA)

http://www.sagenet.org/sage_protocol.htm

This is the web-site of the scientists who developed SAGE. The SAGE software was down-loaded from the web-site after registration.

SAGE software version 3.04 beta

SAGE95, the original software, and SAGE2000, the most recent update, also used.

The SAGE software was down-loaded from the web-site after registration and installed using a .exe file received separately as an email attachment.

SAGE detailed protocol

Versions 1b, November 1995; 1c, September 1997; 1d, November 1999

The detailed SAGE protocols were received as email attachments.

National Centre for Biotechnology Information (NCBI) (Bethesda, MD, USA)

NCBI is an international resource for molecular biology information.

http://www.ncbi.nlm.nih.gov/

NCBI's GenBank® databases: but actually down-loaded from HGMP-RC UK mirror site (see below)

http://www.ncbi.nlm.nih.gov/SAGE/

NCBI's SAGE databases for down-loading and SAGEmap tools for on-line analysis

http://www.ncbi.nlm.nih.gov/UniGene/index.html/

NCBI's UniGene databases and tools for on-line analysis (the "unique gene" database clusters all transcripts corresponding to one gene under one name) http://www.ncbi.nlm.nih.gov/BLAST/

NCBI's BLAST® (Basic Local Alignment Search Tool) set of programs for searching for DNA or protein similarity through all of the available sequence databases, plus alignment programs

http://www.ncbi.nlm.nih.gov/PubMed/

The National Library of Medicine's PubMed database, providing access to over 11 million MEDLINE citations back to the mid-1960's, used here for all literature searches

UK Human Genome Mapping Project Resource Centre (Cambridge, UK)

http://www.hgmp.mrc.ac.uk/About/SEQNET/

UK mirror site of NCBI from where GenBank® databases downloaded

Celera Genomics (Rockville, MD, USA)

http://www.celera.com

USA site of Celera corporation from where the human genomic sequence for foveolin was obtained.

Ensembl (a joint project between EMBL - EBI and the Sanger Institute)

http://www.ensembl.org/

Ensembl is a software system which produces and maintains automatic annotation on eukaryotic genomes. Whole Genome Shotgun sequencing reads generated by the Mouse Sequencing Consortium are available via the Trace Archive at Ensembl, so this is the site from where the mouse genomic sequences were downloaded.

EMBL-European Bioinformatics Institute (EMBL-EBI) (Cambridge, UK)

http://www.ebi.ac.uk

MultAlign multiple sequence alignment (nucleotide)

ClustalW multiple sequence alignment (amino acid)

These programs were used to align and compare sequences of the new gastric gene. The EBI site was used but these programs are widely available.

Mitelman Database of Recurrent Chromosome Aberrations in Cancer

http://cgap.nci.nih.gov/Chromosomes/Mitelman

Pôle Bio-Informatique Lyonnais

http://npsa-pbil.ibcp.fr/NPSA

The Network Protein Sequence Analysis programs at the PBIL provide: primary structure analysis, including physico-chemical profiles; and secondary structure prediction.

PSORT WWW Server

http://psort.nibb.ac.jp/

PSORT is a computer program for the prediction of protein localization sites in cells. Amino acid sequences are analysed by applying general rules for various sequence features of known protein sorting signals. The program then reports the likelihood for the input protein to be localized at each candidate site (Nakai *et al.* 1992).

Ambion Inc (Houston, TX, USA)

http://www.ambion.com/

Ambion company web-site with excellent advice on working with RNA

Qiagen Ltd (Crawley, UK)

http://www.qiagen.com/

Qiagen company web-site with excellent advice on molecular biology in general

2.1.13.2 Software

Microsoft® (Reading, UK)

Microsoft® Access database and Excel spreadsheet programs

Informax Inc (Oxford, UK)

Vector NTI Suite

DNASTAR Inc (Madison, WI, USA)

Lasergene suite containing EditSeq and SeqMan programs

BioRad Laboratories Ltd (Hemel Hempstead, UK)

Molecular Analyst Software

2.1.13.3 SAGEmap libraries used for comparisons with gastric samples

http://www.ncbi.nlm.nih.gov/SAGE/

SAGE libraries from normal glandular epithelial and mesothelial tissues

Breast Br_N

Colon NC1

Ovary HOSE_4

Pancreas HX & 126

Peritoneum PERITO-13

Prostate Chen_normal_prostate

SAGE libraries from adenocarcinomas and mesothelioma

Breast tumor 95-259

Colon tumor TU102

Mesothelioma MESO-12

Ovary tumor OVT7

Pancreas tumor 96-6252

Prostate tumor Chen_Tumor_Pr

2.1.14 Cloning of candidate genes

Invitrogen Life Technologies Ltd (Paisley, UK)

TOPO TA Cloning® Kit with pCR®II-TOPO®

Qiagen Ltd (Crawley, UK)

Qiagen Plasmid Maxi/Midi Kit

2.1.15 Northern blotting and probing of blots

Ambion Inc (Austin, TX, USA)

NorthernMax™ Kit

Strip-EZ™ PCR StripAble™ PCR Probe Synthesis and Removal Kit

BrightStar-Plus Positively Charged Nylon Transfer Membrane

Human Internal Standard Screening Kit, containing a beta-actin template cloned in pTRIPLEscript vector, plus other internal standards and primers

Amersham Pharmacia Biotech UK Ltd (Little Chalfont, UK)

[alpha-^{32}P] dATP (3,000 Ci/mmol, 10 mCi/ml)

[alpha-^{32}P] dCTP (3,000 Ci/mmol, 10 mCi/ml)

Kodak X-OMAT AR-5 Scientific Imaging (X-ray) Film

Rediprime™ II random prime labelling system

Whatman International (Maidstone, UK)

3MM filter paper

2.1.16Immunohistochemistry

2.1.16.1 Primary antibodies

Dako Ltd (Cambridge, UK)

Gastrin	(rabbit anti-human monoclonal antibody)
Lysozyme	(rabbit anti-human polyclonal antibody)
Cytokeratin 8	(mouse anti-human monoclonal antibody)

Lab Vision (UK) Ltd (Neomarkers antibodies) (Newmarket, UK)

pS2	(mouse anti-human monoclonal antibody)
MUC5AC	(mouse anti-human monoclonal antibody)

Novocastra Laboratories Ltd (Newcastle upon Tyne, UK)

Human spasmolytic polypeptide	(mouse anti-human monoclonal antibody)

2.1.16.2 Secondary staining

Vector Laboratories Ltd (Peterborough, UK)

VECTASTAIN Universal Elite ABC-Peroxidase Kit

2.1.17Non-isotopic in situ hybridisation

Amersham Pharmacia Biotech UK Ltd (Little Chalfont, UK)

Dextran sulphate

Fluka BioChemika (Buchs, Germany)

Formamide

New England Biolabs (UK) Ltd (Hitchin, UK)

Kpn I

Not I

Roche Diagnostics Ltd (Lewes, UK)

Anti-Digoxigenin-Alkaline Phosphatase

DIG RNA Labeling Kit (Sp6/T7)

NBT/BCIP stock solution (nitrobluetetrazolium/bromochloroindolylphosphate)

Sigma (Dorset, UK)

Alcian blue

Denhardt's solution

Glycergel (glycerol gelatin)

Haematoxylin

Levimasole

Proteinase K

Salmon sperm DNA

Triton X-100™

ISH pre-hybridisation buffer

Make up 100 ml using: 10 ml 20X SSC, 50 ml 100% Formamide and 40 ml DEPC-treated dH_2O

ISH hybridisation buffer

Make up 20 ml using: 40 µl 5 M Tris (pH 7.5), 2.5 ml 100X Denhardts, 1.3 ml 20X SSC, 500 µl 20% SDS, 10 ml 100 % Formamide, 4 ml 50% Dextran Sulphate, 500 µl salmon sperm DNA (10 mg/ml) and 11.6 ml DEPC-treated dH_2O

Dig (1) buffer

100 mM Tris base, 100 mM Tris HCl and 150 mM NaCl (solution pH 7.6)

Dig (3) buffer

100 mM Tris base (pH 9.5), 100 mM NaCl and 50mM Magnesium Chloride

2.1.18 Fluorescent in situ hybridisation (FISH)

Qbiogene (Appligene Oncor) (Middlesex, UK)

Chromosome 2 paint

Roche Diagnostics Ltd (Lewes, UK)

4',6-diamidino-2-phenylindole (DAPI)

Sigma (Dorset, UK)

NP-40 (Tergitol, a non-ionic surfactant)

Pepsin

Propidium iodide (PI)

RNase (ribonuclease)

Streck Laboratories Inc (La Vista, NE, USA)

Streck Tissue Fixative

Vector Laboratories Ltd (Peterborough, UK)

VECTASHIELD® Mounting Medium

Vysis Inc (Downers Grove, IL, USA)

Nick Translation Kit with SpectrumGreen dUTP

2.1.19Probing of genomic library

UK Human Genome Mapping Project Resource Centre (Cambridge, UK)

Human Genomic PAC Library RPCI1 supplied as high density gridded filters for screening by hybridisation

Human Genomic PAC Library RPCI1, Clone 291-N4

2.2 Methods

2.2.1 Collection and preparation of tissue and cell samples

2.2.1.1 Primary tissues for SAGE, Northern blotting and immunohistochemistry

Resection specimens of oesophagus, stomach, duodenum and colon were collected from the operating theatre immediately after surgery and taken to the Pathology Department, North Glasgow Hospitals University NHS Trust. There, I (a pathologist) examined and dissected the specimens. Diagnostic tissue samples were taken for routine histopathology from the tumors, resection margins, normal mucosa and lymph nodes. These diagnostic tissue blocks were fixed in formalin, processed through to paraffin wax, sectioned and stained with Haematoxylin and Eosin, and, where necessary, mucin stains (here, Alcian Blue). This was performed by pathology technical staff, according to standard operating procedures. I then examined the microscopic slides and reported the cases.

For this research project, tissue samples were taken from immediately adjacent to the blocks selected for diagnosis, diced into 5mm^3 pieces, flash-frozen in liquid nitrogen and stored at -70°C. Tissue collection was in accordance with contemporary ethical practice. A range of primary tumours and normal mucosal tissues was collected. The gastric adenocarcinomas are detailed in Table 2.1 (Section 2.1.5.1), of which Tumours 1 and 2 were used for SAGE, as was a sample of normal (non-neoplastic) gastric antral mucosa taken from a Whipple's resection from a patient with pancreatic (but not gastric) adenocarcinoma. The non-neoplastic mucosa showed a chronic gastritis without dysplasia.

For SAGE, RT-PCR and Northern blotting, RNA was prepared from the frozen samples. For immunohistochemistry and *in situ* hybridisation, the formalin-fixed, paraffin-embedded tissue was used.

2.2.1.2 Culture of cell lines

Cells were cultured, subcultured and stored according to the suppliers' instructions. AGS cells (Barranco *et al.* 1983) were cultured in Ham's F12 medium supplemented with 10% (v/v) Foetal Bovine Serum (FBS). KATO-III cells (Sekiguchi *et al.* 1978) and MKN-45 cells (Naito *et al.* 1984) were cultured in RPMI 1640 medium supplemented with 20% FBS. OE19 cells (Rockett *et al.* 1997) were cultured in RPMI 1640 medium supplemented with 10% FBS. All media were also supplemented with 2mM L-Glutamine, 500 units/ml Penicillin and 500 µg/ml Streptomycin. All cell lines were maintained in a humid 37°C / 5% CO_2 incubator.

For Northern blotting, RNA was prepared directly from the cells as described in Section 2.2.2.2. For immunohistochemistry and *in situ* hybridisation, the cells were harvested according to the suppliers' instructions, fixed in formalin, then pelleted in agar by brief centrifugation. The agar pellet of cells was then processed through to paraffin wax as for the tissue samples.

2.2.1.3 Archival tissues for profiling of the new gastric gene

Archival blocks were selected from the files in the Pathology Department. The human tissues examined included a wide range of normal tissues and their corresponding tumours, principally adenocarcinomas and their originating glandular epithelia (see Sections 2.1.5.1 and 2.1.5.2). The tissues were fixed in formalin, processed routinely, embedded in paraffin and sectioned in the conventional manner. Tissue samples of mouse and rat stomach and of bovine abomasum (the fourth stomach, anatomically closest to the human) were also examined.

2.2.2 General molecular biological manipulation of RNA and DNA

2.2.2.1 Maintenance of RNase-free working environment

Many methods described in this thesis use RNA as their starting material, so that my experimental results depended on the quality and quantity of input RNA (see Section 3.1.2.1). However, RNA is highly labile and subject to degradation by ribonucleases (RNases). The main source of RNase contamination is endogenous RNases which must be quickly inactivated during cell lysis, usually with protein denaturants such as phenol, chloroform and chaotropic salts, which are contained in TRIzOL® Reagent (Section 2.2.2.2). Exogenous RNases may also cause problems but were avoided through various measures. Disposable gloves were worn and separate pipettes were used for RNA. Water for solutions was treated by adding diethylpyrocarbonate (DEPC) to a final concentration of 0.1% (v/v), then the solution was mixed, left overnight in a fume hood and autoclaved. Where possible, RNase-free sterile plasticware was used. Glassware was baked at 150°C for 4 h or rinsed with 0.5 M NaOH. Other equipment, including pipettes, mortar and pestle, metalware and electrophoresis equipment, and bench surfaces were treated with Ambion's RNaseZap™ (RNase Decontamination Solution).

2.2.2.2 Purification of total RNA

TRIzOL® Reagent was used to isolate total RNA from frozen tissues and cultured cells, according to manufacturer's instructions.

For frozen tissue samples, 2 ml TRIZOL® was used per 100 mg of tissue. For SAGE, 500-1,000 mg of tissue was used (with a 50 ml conical tube). For Northern blotting and other verification procedures, 100 mg of tissue was used (with two 1.5 ml microcentrifuge tubes). The tissue pieces were not allowed to thaw during manipulation but were weighed, then crushed into small fragments in liquid nitrogen using a mortar and pestle, and finally powdered using a bead mill in the Mikrodismembrator. TRIZOL® was poured on the frozen tissue powder and mixed in during gradual defrosting. For adherent cell lines, 1 ml TRIZOL® was used per 10 cm^2 of the area of the culture flask (with a 50 ml conical tube). For spherical (in suspension) cell lines, the cells were pelleted by centrifugation then 1 ml TRIZOL® was used per 5-10 x 10^6 cells. The TRIZOL® was poured onto the cells and the lysate passed several times through a pipette. With the larger volumes of tissue or cell lysate, a glass-Teflon® homogeniser was used. All samples were further homogenised by passing through a needle and syringe.

Homogenised samples were then incubated for 5 min at RT. 0.2 ml chloroform was added per 1 ml of TRIZOL®. The tubes were shaken for 15 s, incubated at RT for 3 min then centrifuged. 50 ml conical tubes were spun in the Beckman centrifuge at 4,000 rpm at 4°C for 30 min. Microcentrifuge tubes were spun at full speed at 4°C for 15 min. The colourless upper aqueous phase was transferred to a new tube. 0.5 ml isopropanol was added per 1 ml of TRIZOL® used initially. The samples were incubated at RT for 10 min then centrifuged again, causing the RNA precipitate to pellet. The supernatant was removed and the pellet washed with 70% ethanol then air-dried for 5 min and resuspended in dH_2O.

The yield of total RNA was quantitated using a spectrophotometer (Section 2.2.2.9) and its quality assessed by electrophoresis on a 1% (w/v) agarose gel (Section 2.2.2.6). Two strong, distinct bands of ribosomal RNA should be visible at 4718 and 1847 nucleotides, representing the 28S and 18S subunits, but significant smearing indicates RNA degradation. The RNA samples were then stored at -70°C.

2.2.2.3 Phenol/chloroform extraction of DNA

Phenol extraction is used to remove protein from solutions of nucleic acids. An equal volume of phenol/chloroform/isoamyl alcohol (25:24:1, v/v/v) (P/C) was added to the sample, mixed well by vortexing and spun in a microcentrifuge at full speed at RT for 5 min, to separate the phases. Nucleic acids partition into the aqueous (top) phase, which was transferred to a fresh microcentrifuge tube, without disturbing the lower layer or interface.

2.2.2.4 Ethanol precipitation of DNA or RNA

Ethanol precipitation is used to purify and reduce the volume of solutions of nucleic acids. A one-third volume of 10 M ammonium acetate and 3 volumes of 100% ethanol were added to the sample. In SAGE, 3 µl glycogen was also added, as a carrier to enhance precipitation. The sample was mixed well by vortexing and allowed to precipitate on dry ice (or at -20°C) for at least 15 min, then centrifuged at full speed, preferably at 4°C, for 15 min. The pellet was washed with 70% ethanol, re-spun then dried and resuspended in dH_2O or LoTE.

2.2.2.5 Purification of DNA using QIAquick columns

DNA purification by P/C extraction and ethanol precipitation has been described above but Qiagen's QIAquick® Spin Columns provide an alternative which at some stages save time and possibly provide purer DNA samples. The purification columns of the QIAquick® PCR Kit may be either spun at full speed in a microcentrifuge or extracted with a vacuum manifold (for large-scale work using the QIAquick® 8 Kit and QIAvac 6S). The DNA sample was mixed with five volumes of Buffer PB and loaded onto the QIAquick column. The column was then spun (or subjected to vacuum). The column was washed with 0.7 ml Buffer PE and spun. A further spin or vacuum was applied to dry the column. To elute the DNA, 30-100 µl Buffer EB was placed on the membrane and a final spin or vacuum applied.

2.2.2.6 Agarose gel electrophoresis

DNA produced by PCR and RNA were analysed using non-denaturing agarose gel electrophoresis which separates nucleic acids according to size. Electrophoresis grade agarose at between 0.8% and 2% (w/v) was added to 0.5X TBE buffer and heated in a microwave oven to dissolve the agarose. The molten gel was allowed to cool to approximately 60°C then poured into a casting tray and left to set. The gel was run in a gel tank in 0.5X TBE buffer at 100-200 V. For loading, the samples were mixed with a one-fifth volume of gel loading buffer. DNA molecular weight standards, usually 100 bp ladder and λHindIII, were run alongside the samples. To visualise the nucleic acids, ethidium bromide (0.5 µg/ml) was added to the molten gel, then after electrophoresis the gel was examined under UV transillumination, using the BioRad Gel Doc 1000 gel documentation system and Molecular Analyst software. This system also enables gel photography.

2.2.2.7 Polyacrylamide gel electrophoresis (PAGE)

During SAGE, the PCR products and ditags were isolated by 12% PAGE and the ligated concatemers were separated by 8% PAGE. 12% PAGE used 14 ml of 40% polyacrylamide (19:1 acrylamide:bis) and 31.3 ml dH_20. 8% PAGE required 9.3 ml of 40% Polyacrylamide (37.5:1 acrylamide:bis) and 36 ml dH_20. To either mix, 930 μl 50X Tris Acetate Buffer, 470 μl 10% APS and 30 μl TEMED were added. The mixture was poured into the vertical gel apparatus and left to polymerise for 30 min. The gel was run in 1X TAE Buffer at 130-160 V for 2-3 hr, as described in the text. After electrophoresis, the gel was stained with SYBR® Green I, in 150 ml of a 1:10,000 dilution of the stock solution in a light-resistant container at RT for 30 min. The gel was visualized with UV transillumination, using the BioRad Gel Doc 1000 gel documentation system and Molecular Analyst software.

2.2.2.8 Polymerase chain reaction

Except where otherwise specified, most PCR amplifications were performed using Qiagen's *Taq* PCR core kit as follows. Each 100 μl reaction volume contained: 60.5 μl dH_2O, 10 μl 10X Qiagen PCR buffer, 5 μl 25 mM $MgCl_2$, 20 μl 5X Q-Solution, 2 μl 10mM dNTPs, 0.5 μl of each of two oligonucleotide primers, 0.5 μl *Taq* DNA Polymerase and 1 μl input DNA. Where necessary, the volumes were scaled downwards. PCR was performed using a hot-lid or with a drop of oil covering the surface of the reaction mix. The most commonly used PCR machines were Hybaid's Touchdown and Omnigene Thermal Cyclers with hot-lids. The cycling parameters were: 94.5°C for 1.5 min; 30 cycles of 94.5°C for 30 s, 52°C for 1 min and 72°C for 1 min; then 72°C for 5 min. Occasional changes to the annealing temperature or concentration of magnesium ions were required for optimal results. Where problems arose in getting the PCR to work initially or in obtaining specificity, Qiagen's HotStarTaq™ DNA Polymerase Kit was used. Stategies used to prevent cross-contamination included treatment of PCR equipment with Ambion's DNAZap™ PCR DNA Degradation Solution, or with UV light in the UV Stratalinker 2400; and treatment of PCR solution components with UV light.

For SAGE, the oligonucleotide primers used were as described in the Johns Hopkins' protocol (Section 2.1.8.1). For RT-PCR, the primers were designed personally using the down-loaded cDNA/mRNA sequences of the candidate genes and Vector NTI or PrimerDesign in DNAStar software (Section 2.1.8.2).

2.2.2.9 Quantitation of RNA or DNA concentration

DNA was quantified by spectrophotometric determination of its UV light absorbency. An aliquot of the sample, or a dilution of it in dH_2O, was placed in a capillary tube and its absorption measured at the ultraviolet light wavelengths of 260 nm and 280 nm in a quartz cuvette, using dH_2O as a blank. The concentration of nucleic acid in µg/ml was therefore calculated by multiplying the A260 x dilution factor x 40 (for RNA) or 50 (for DNA). Nucleic acid purity was assessed using the ratio of A260/A280 readings which should fall between 1.8 and 2.0. The GeneQuant RNA/DNA Calculator Spectrophotometer which was used in fact directly displayed the DNA and RNA absorbances and concentrations plus the A260/A280 ratios.

2.2.2.10 Sequencing

The DNA to be sequenced was made up to a volume of 5 µl (30-90 ng PCR product DNA, 300-600 ng BAC DNA) and mixed with 1 µl sequencing primer (approximately 3.2 pmoles) and 4 µl reaction pre-mix from the ABI PRISM® BigDye™ Terminator Cycle Sequencing Ready Reaction Kit.. PCR was performed on the PTC-100™ Programmable Thermal Controller. The cycling parameters were: preheat to 96°C; 25 cycles of 96°C for 10 s, 50°C for 5 s and 60°C for 4 min; then hold at 4°C. The samples were then ethanol precipitated, basically according to the previous protocol, but with precipitation and centrifugation at RT, to minimise precipitation of labelled free nucleotides which impair the sequencing gel electrophoresis.

The remainder of the sequencing was kindly performed by the local Beatson Laboratories sequencing service staff, headed by Robert McFarlane. The samples were resuspended in loading buffer, heated and quickly chilled, loaded onto the gel and run overnight on an ABI Prism Automated Sequencer (Models 373 and 377).

2.2.3 Serial Analysis of Gene Expression (SAGE)

2.2.3.1 SAGE laboratory procedures

The SAGE method described is based on the protocols from Johns Hopkins University, versions 1b, November 1995, and 1c, September 1997. My own modifications are described as such in the text.

2.2.3.1.1 Preparation of mRNA from cells or tissues

Total RNA was prepared using TRIzol® Reagent (Section 2.2.2.2). Polyadenylated (poly(A), messenger) RNA was purified from total RNA using Ambion's Poly(A)Pure™

Kit, according to manufacturer's instructions. The Johns Hopkins protocol stipulated the need for 5 μg mRNA which is broadly equivalent to 500 μg total RNA, 500 mg tissue or 5 x 10^7 cultured cells (technical notes, web-sites of Ambion and Qiagen). I used 500-1,000 mg of tissue as starting material.

Total RNA was mixed with 0.1 volumes of 5 M NaCl and transferred to a 15 ml conical tube. Binding Buffer was added to raise the volume to 4 ml. The RNA was heated at 65°C for 5 min then quickly chilled on ice for 1 min. One vial of Oligo(dT) Cellulose was added and mixed then incubated at RT for 60 min with gentle rocking to enable the poly(A) mRNA to bind. The cellulose was pelleted by centrifugation at 4,000 rpm in the table-top Beckman GS-6R centrifuge at RT for 3 min. The supernatant was discarded. Through serial incubations and centrifugations, the Oligo(dT) Cellulose was washed three times with 10 ml Binding Buffer and three times with Wash Buffer. The Oligo(dT) Cellulose was transferred to a filter spin column in a 2 ml tube where it was washed three times with 0.5 ml Wash Buffer with intervening spins in a microcentrifuge at 4,000 rpm. The mRNA was eluted with two aliquots of 200 μl Elution Buffer, pre-warmed to 70°C. The 400 μl flow-through, containing the mRNA, was ethanol precipitated and resuspended in 10 μl LoTE.

During the initial set up of SAGE, many alternative strategies for preparing mRNA were tested (Section 3.1.2.1) but the method described was preferred.

2.2.3.1.2 cDNA synthesis

Double-stranded cDNA was prepared using Invitrogen's cDNA Synthesis System, according to manufacturer's instructions. For first strand synthesis, the 10 μl of mRNA was mixed with 10 μl 5X First Strand Buffer, 2.5 μl 10 mM dNTP mix, 5 μl 0.1 M dithiothreitol (DTT), 7.5 μl (2.5 μg) of the separately purchased biotinylated oligo dT (instead of that supplied with the kit), 12.5 μl dH_2O and 2.5 μl Moloney Murine Leukaemia Virus (M-MLV) reverse transcriptase then incubated at 37°C for 1 h.

For second strand synthesis, 289.5 μl dH_2O, 7.5 μl 10 mM dNTP mix, 40 μl 10X Second Strand Buffer, 10 μl *E. coli* DNA Polymerase I, 1.75 μl *E. coli* RNase H and 1.25 μl *E. coli* DNA Ligase were added directly to the first strand reaction and incubated at 16°C for 2 h. The DNA was phenol/chloroform (P/C) extracted, ethanol precipitated and resuspended in 20 μl LoTE.

In initial experiments, a radioactive ^{32}P tracer was used in aliquots of the first and second strand reaction mixes to determine the yield of cDNA by precipitation with trichloroacetic acid (TCA), then scintillation counting, according to manufacturer's instructions. Since the yields were appropriate, this tracer step was discontinued.

2.2.3.1.3 Cleavage of biotinylated cDNA with anchoring enzyme NlaIII to create CATG sticky-end

10 µl (half) of the biotinylated cDNA was mixed with 74 µl LoTE, 10 µl 10X NEBuffer 4, 1 µl 100X BSA and 5 µl *Nla* III and incubated at 37°C for 1 h, then P/C extracted, ethanol precipitated and resuspended in 20 µl LoTE.

2.2.3.1.4 Binding of biotinylated cDNA to magnetic beads

100 µl Streptavidin Dynabead slurry was added to each of two 1.5 ml microcentrifuge tubes. The beads were immobilised with the Dynal magnet and the supernatant removed. The beads were washed with 200 µl 1X B+W buffer. 100 µl 2X B+W buffer, 90 µl dH_2O and 10 µl cleaved biotinylated cDNA were added to both tubes which were incubated with gentle mixing at RT for 15 min. The beads were then washed three times with 200 µl 1X B+W buffer and once with 200 µl LoTE.

2.2.3.1.5 Ligating CATG sticky-ended linkers (kinased and annealed in advance) to bound cDNA

The linker oligonucleotides were enzymatically kinased and annealed in advance. Two tubes were set up, one each for linker pairs 1 and 2. 9 µl of Linker B (either 1B or 2B) was mixed with 8 µl LoTE, 2 µl 10X Ligase Buffer (which contains ATP and in which the Kinase enzyme has 100% activity) and 1 µl T4 Polynucleotide Kinase. The reactions were incubated at 37°C for 30 min then heat inactivated at 65°C for 10 min. 9 µl Linker 1A was added to kinased Linker 1B, and similarly for Linkers 2. The linkers were annealed by heating to 95°C for 2 min, then allowing to cool to RT over 15 min.

One of the two tubes containing washed magnetic beads was used for linker 1 and the other for linker 2. The beads were immobilised and the LoTE removed. 29 µl LoTE, 5 µl annealed linker 1 or 2, and 4 µl 10X Ligase Buffer were added to each tube, mixed, heated at 50°C for 2 min then allowed to cool to RT over 15 min. 2 µl T4 DNA Ligase was added and the reaction incubated at 16°C for 2 h with intermittent gentle mixing. After ligation, the beads were washed four times with 200 µl 1X B+W buffer. and twice with 200 µl 1X NEBuffer 4.

2.2.3.1.6 Creation of cDNA tags and their release from magnetic beads using tagging enzyme BsmF I

The buffer was removed from the beads. 87 µl LoTE, 10 µl 10X NEBuffer 4 and 1 µl 100X BSA were added. The tubes were pre-incubated at 65°C for 2 min then 2 µl *BsmF* I was added and incubated at 65°C for 1 h with intermittent gentle mixing. The beads were immobilised and, this time, the supernatants were collected and transferred to two new tubes. The beads were washed with 100 µl 1X NEBuffer 4 which was then added to the previous supernatants. The beads were discarded. The supernatants were P/C extracted, ethanol precipitated and resuspended in 10 µl LoTE.

2.2.3.1.7 Blunt-ending BsmF I-created sticky-ends of released cDNA tags

To each of the two tubes (from Linkers 1 and 2) containing released cDNA tags, 31 µl dH_2O, 5 µl 10X EcoPol Buffer, 0.5 µl 100X BSA, 2.5 µl 10 mM dNTPs and 1 µl DNA Polymerase I Large Fragment (Klenow) were added. The reactions were incubated at 37°C for 30 min, then pooled in one tube, P/C extracted, ethanol precipitated and resuspended in 12 µl LoTE.

2.2.3.1.8 Ligating blunt-ended tags to form 102 bp ditags

Two 0.2 ml tubes were set up, of which one was for the ditag ligation reaction. The other was a negative ligation control and was set up first with the purpose of excluding or identifying cross-contamination at the next, PCR, step. To both tubes, 4 µl blunt-ended tags, 0.8 µl dH_2O and 0.6 µl 10X Ligase Buffer were added. 0.6 µl dH_2O was added to the negative control tube. Then 0.6 µl T4 DNA Ligase was added to the ditag reaction. The reaction mixes were covered with a drop of mineral oil to prevent evaporation of the small reaction volume and incubated at 16°C overnight.

2.2.3.1.9 PCR amplification of 102 bp ditags

This PCR step aims to produce sufficient 102 bp ditag DNA for subsequent isolation and concatemerisation of 26 bp ditags, but itself was often problematic.

After overnight incubation, 14 µl LoTE was added to the ligation reaction and mixed. 1 µl of this mixture was removed and diluted 100-fold with LoTE. 1 µl of this dilution was then used in a 50 or 100 µl PCR reaction with SAGE Primers 1 and 2. To prevent PCR cross-contamination, the two negative control reactions (no template and no ligase) were set up first.

The original Johns Hopkins protocol described 50 µl PCR reactions containing: 30.5 µl dH_2O, 5 µl 10X SAGE PCR buffer, 3 µl DMSO, 7.5 µl 10mM dNTPs, 1 µl of each of SAGE Primers 1 and 2, and 1 µl PLATINUM® *Taq* DNA Polymerase. The cycling parameters, optimized for a Hybaid thermal cycler, were: 94°C for 1 min; 26-30 cycles of 94°C for 30 s, 55°C for 1 min and 70°C for 1 min; then 70°C for 5 min. The PCR reactions required to be optimised with different template dilutions (1/50, 1/100 or 1/200 per reaction).

I found that the Johns Hopkins PCR conditions were not robust (Section 3.1.2.2). My own modification used Qiagen's HotStarTaq™ DNA Polymerase Kit, according to manufacturer's instructions, which routinely worked well with 1 µl, or often less, of the 1/100 dilution with no need for further adjustment of template dilution. Each 100 µl reaction contained: 60.5 µl dH_2O, 10 µl 10X Qiagen PCR buffer, 5 µl 25 mM $MgCl_2$, 20 µl 5X Q-Solution, 2 µl 10mM dNTPs, 0.5 µl of each of SAGE Primers 1 and 2, and 0.5 µl HotStarTaq™ DNA Polymerase. The cycling parameters were: 94.5°C for 15 min; 26-30 cycles of 94.5°C for 30 s, 56°C for 1 min and 72°C for 1 min; then 72°C for 5 min.

The Johns Hopkins protocol recommended optimisation of the cycle numbers, between 26 and 30. I found that 29 cycles usually produced good results. More than 30 cycles resulted in high molecular weight smearing with less of the desired product, as found previously (Bell *et al.* 1991).

After PCR, 10 µl of each reaction was loaded on a 12% polyacrylamide gel with a DNA ladder (Life Technology's 10 bp ladder or Roche's DNA Molecular Weight Markers V and VIII). The gel was run at 160 V for 2.5 hr then stained and examined (Section 2.1.10). The amplified ditags should produce a 102 bp band. Background bands are common: the brightest runs at 80 bp and contains amplified ligated linkers without tags. The negative controls should contain no product.

After optimisation, large-scale PCR was performed by preparing then distributing a master-mix into three 96-well PCR plates with 100 µl per well. After PCR, the reactions were pooled into 50 ml conical tubes, then P/C extracted, ethanol precipitated and resuspended in a total of 250 µl LoTE.

2.2.3.1.10 Isolation of 102 bp ditags by gel-purification

The pooled PCR products were loaded on three 12% polyacrylamide gels which were run and stained as before. The 102 bp band of amplified ditags was cut out. The pieces of gel were fragmented using 0.5 ml microcentrifuge tubes which were pierced through their base with a needle then inserted into a 2.0 ml tube. Three tubes were used per gel. The excised gel bands were placed in the 0.5 ml tubes and spun at full speed for 2 min. The DNA was eluted from the gel fragments by adding 250 µl LoTE and 50 µl 10 M ammonium acetate to each 2.0 ml tube. The tubes were vortexed then incubated at 65°C for 2 h. The contents of each tube was transferred to two Spin-X filter microcentrifuge tubes which were spun at full speed for 5 min. The eluted 102 bp ditag DNA was pooled, ethanol precipitated and resuspended in a total of 100 µl LoTE.

2.2.3.1.11 Isolation of 26 bp ditags by NlaIII digestion and gel-purification

To the pooled PCR products, 58 µl LoTE, 20 µl 10X NEBuffer 4, 2 µl 100X BSA and 20 µl *Nla* III were added. The reaction was incubated at 37°C for 1 h, then P/C extracted, ethanol precipitated and resuspended in 15 µl LoTE. The DNA was loaded on two lanes of a 12% polyacrylamide gel which was run at 130 V for 2.5 hr then stained. The ditag band running at 22-26 bp was excised and eluted as before, this time incubating at 37°C not 65°C. The 26 bp ditag DNA was ethanol precipitated and resuspended in 6.4 µl LoTE.

2.2.3.1.12 Ligation of sticky-ended 26 bp ditags to form concatemers then gel-purification of concatemers

To the purified ditags, 0.8 µl 10X Ligase buffer and 0.8 µl T4 DNA Ligase were added and incubated at 16°C overnight. Loading buffer was added directly to the ligation reaction which was loaded in one lane of an 8% polyacrylamide gel with a 100 bp ladder. The gel was run at 130 V for 3 hr then stained. The DNA smear over 500 bp in size was cut out and eluted as before, this time incubating at 65°C. The concatemer DNA was ethanol precipitated and resuspended in 6 µl LoTE.

During the initial set up of SAGE, different conditions for the ligation reaction and alternative strategies for isolating the longer concatemers were tested but the (original Johns Hopkins) method described proved to be as good as any others.

2.2.3.1.13 Cloning concatemers

The concatemers were cloned using the Zero Background™ Cloning Kit. The pZErO®-1 vector in this kit contains a lethal gene which is disrupted by DNA insertion, so that only positive recombinants should grow (this is the theory, but in practice some colonies did lack inserts). To linearise the vector, 1 µl pZErO®-1 (1 µg/µl) was mixed with 7 µl dH_2O, 1 µl NEBuffer 2 and 1 µl *Sph* I. The restriction enzyme digest was incubated at 37°C for 30 min. The vector DNA was P/C extracted and ethanol precipitated then resuspended in 30 µl LoTE. 1 µl *Sph* I-linearised pZErO® was mixed with the 6 µl of purified concatemers, 1 µl 10X Ligase buffer and 1 µl T4 DNA Ligase, then incubated at 16°C for 2 hr. Two control reactions were also set up: a no insert control, omitting concatemers; and a no ligase control, omitting both concatemers and ligase. The reactions were P/C extracted, ethanol precipitated and resuspended in 3 µl LoTE.

Before transformation, Low Salt LB agar plates containing 50 µg/ml Zeocin and 1 mM IPTG were prepared, according to manufacturer's instructions. Usually 1.5 litres of medium were made up and poured into around 30 plates of 140 mm diameter. 1 µl of the three ligation and control reactions was placed in 0.5 ml microcentrifuges tube and chilled on wet ice, along with the electroporation cuvettes. One vial of ELECTROMAX^{TM} DH10B^{TM} *E. coli* Cells was thawed on wet ice, then mixed by tapping gently. Taking one sample at a time, 20 µl of cells was added to the chilled 1 µl reaction, mixed gently, then transferred to a chilled electroporation cuvette. The cuvette was placed in the chamber of the Cell Shock Electroporator and the electrical pulse discharged. The cuvette was removed and 1 ml SOC medium at RT was immediately added and the mixture transferred to a 14 ml Falcon snap-cap round-bottomed tube. This procedure was repeated for all ligation or control reactions. The tubes were incubated with shaking at 225 rpm at 37°C for 60 min.

The cultures were then diluted with 2 ml SOC medium. The transformation mix was spread onto the Low Salt LB-Zeocin plates at 100 µl per plate. All of the test sample was plated but only one plate was used for each of the control reactions. The liquid was left to absorb then the plates were inverted and incubated at 37°C overnight. The plates were then removed from the incubator and examined. Zeocin-resistant transformants were picked for further analysis. All plates were stored at 4°C, until the inserts had been checked, for later use in large-scale sequencing.

With the Zero Background^{TM} Cloning Kit, although in theory only positive recombinant colonies should grow, in practice many experiments resulted in either all or a majority of colonies lacking inserts. I therefore used similar methods to those desribed above with an alternative vector, pGEMR-3Zf(+) from ABI PRISM® BigDye^{TM} Terminator Cycle Sequencing Ready Reaction Kit. This vector uses ampicillin-resistance and IPTG/X-Gal blue-white selection but the results overall were similar (see Section 3.1.2.4.3).

2.2.3.1.14 Screening of transformants by PCR to identify long concatemer inserts

The Zeocin-resistant transformants were then used in a PCR reaction with vector-specific primers to determine the size of the DNA insert size in each bacterial colony. Initially 0.5 ml tubes were used but, for later large-scale screening, 96-well PCR plates were more efficient. The initial 25 µl reaction volume was later reduced to 16 µl.

The original Johns Hopkins protocol described 25 µl PCR reactions containing: 2.5 µl 10X SAGE PCR buffer; 1.25 µl DMSO; 1.25 µl 10mM dNTPs; 0.5 µl of each of M13 Forward and Reverse Primers; 19 µl dH_2O; and 0.2 µl PLATINUM® *Taq* DNA Polymerase. The cycling parameters were: 95°C for 2 min; 25 cycles of 95°C for 30 s, 56°C for 1 min and 70°C for 30 s; then 70°C for 5 min.

However, I found that the Johns Hopkins PCR conditions were not robust and instead used Qiagen's *Taq* PCR core kit. Each 100 µl reaction volume contained: 61.5 µl dH_2O, 10 µl 10X Qiagen PCR buffer, 5 µl 25 mM $MgCl_2$, 20 µl 5X Q-Solution, 2 µl 10mM dNTPs, 0.5 µl of each of M13 Forward and Reverse Primers, and 0.5 µl *Taq* DNA Polymerase. In fact, the reaction volume was usually 16 µl. The cycling parameters were: 94.5°C for 1.5 min; 30 cycles of 94.5°C for 30 s, 52°C for 1 min and 72°C for 1 min; then 72°C for 5 min.

Each colony was touched with a new pipette tip which was then dipped into the reaction mix and shaken. This was repeated as necessary then PCR performed. 5 µl of each reaction was run on a 2% agarose gel with a 100 bp ladder.

2.2.3.1.15 Sequencing of SAGE concatemer inserts

PCR products of 500 bp in size or over were selected for sequencing, since these should contain at least 15 tags (226 bp of flanking pZErO®-1 vector plus 12-13 bp per tag). Before sequencing, the PCR product was purified, partly to remove the primers, since M13 Forward was used again in the sequencing PCR reaction. Individual P/C extraction and ethanol (or isopropanol) precipitation was one option and indeed was used in the original Johns Hopkins protocol. However, Qiagen's QIAquick® PCR Purification Kits were more efficient for large-scale work, for which I used the QIAquick® 8 PCR Purification Kit along with the QIAvac 6S (Section 2.2.2.5).

The PCR products were then sequenced using the BigDye Kit and M13 Forward Primer with one-half to one-tenth of the purified PCR product per sequencing reaction, as previously described (Section 2.2.2.10).

2.2.3.2 SAGE bioinformatics

2.2.3.2.1 Analysis of SAGE sequence files

Within the concatemer sequences, the linked ditags of approximately 26 bp are separated by CATG, which is the recognition site of *Nla* III. The Johns Hopkins' SAGE software uses the CATG sequence to identify and extract the ditags which are then halved into individual tags. The software then quantifies the number of times the tag occurs within a given population of clone inserts and creates a report of the abundance of each tag. The report can be linked to gene databases for identification of the gene(s) corresponding to the tags and can also be used to compare different SAGE libraries.

The Johns Hopkins' SAGE software was down-loaded, installed and used according to the instructions provided, in combination with NCBI's Genbank® databases, and Microsoft® Access and Excel. Within the SAGE software, a new SAGE project was first created which assembled the database files necessary to keep track of SAGE project data and which recorded information such as the Anchoring Enzyme and tag length. Before allowing the computer to analyze the SAGE data, a few sequences were checked by eye to ensure that they gave the expected intermittent CATG pattern. The sequence data of clones containing concatenated SAGE DiTags was then entered. The SAGE software extracted tag data from the sequence files and added them to the opened project. After checking for duplicate DiTags, individual tags were identified, assigned a Tag Number, and recorded. Then, using the Report command, a list of tags sorted by their frequencies could be generated. Certain tags, such as those derived from the oligonucleotide linkers, were usually excluded from this analysis (see Section 3.1.3 for further details).

The SAGE report is usually saved, as well as in text file format, as an MS Access database file, which allows more complex analysis and searching of the SAGE data.

Potential matching transcripts are identified by linking the report to a SAGE Gene Database. The Genbank genetic databases were downloaded by anonymous File Transfer Protocol (FTP) from SEQNET (the UK academic communities SEQuence NET work computer for molecular biologists, at HGMP Resource Centre, Cambridge). Since the Genbank databases are updated every two months, they were downloaded at these intervals and used to build new databases each time. They were then converted to the appropriate format on a local unix drive and entered into SAGE program gene databases (see Section 3.1.3 for further details). In addition, the National Center for Biotechnology Information (NCBI)'s on-line bioinformatics tools, SAGEmap and Unigene were extremely useful for further analysis of transcripts (Lal *et al.* 1999; Wheeler *et al.* 2001).

2.2.3.2.2 Comparison of SAGE libraries

Other functions within the SAGE program enable the comparison of multiple SAGE projects and generates an incidence sorted list of the tags present in currently opened projects. Statistical analysis is possible, with Significance Calculation which calculates the relative likelihood that a difference would be seen by chance. Performing similar analysis on an entire project the expected differences can be modelled and used to convert relative likelihoods to approximate absolute likelihoods.

To identify genes differentially expressed between gastric carcinoma and normal stomach, the libraries were therefore compared using SAGE software and Microsoft™ Access and Excel programs. Differential expression has been defined in previous SAGE studies by various criteria, including: three-fold, five-fold or ten-fold differences in tag ratio; p-values below 0.05, 0.01 or 0.001; and combinations thereof (Velculescu *et al.* 1995; Zhang, L. *et al.* 1997; Hibi *et al.* 1998; Lal *et al.* 1999; Nacht *et al.* 1999; Velculescu *et al.* 1999; Hough *et al.* 2000; Waghray *et al.* 2001). The definition used here was a difference of five-fold or more combined with a p-value of 0.01 or less.

Genes expressed specifically in the stomach were identified by digital comparison with other glandular epithelial and mesothelial tissues. One normal and one tumor SAGE library from each of breast, colon, ovary, pancreas, prostate and mesothelium were downloaded from the SAGEmap web-site (see Section 2.1.13.3) (Lal *et al.* 1999). Normal stomach was compared pair-wise with the normal libraries, and the pooled gastric carcinomas were compared with the tumors. The criteria for significance were those defined in the last paragraph. Genes which appeared to be expressed specifically in the stomach on the basis of these six comparisons were then checked on-line against the many other normal and tumor libraries in SAGEmap (Lal *et al.* 1999; Wheeler *et al.* 2001).

2.2.4 Validation of gastric SAGE results

To validate and expand the SAGE profiles, selected transcripts were studied in a wider panel of 19 gastro-intestinal tumour and normal tissues and cell lines. Where the genes had been minimally characterised in the stomach, the method of choice was Northern blotting for mRNA. Prior to Northern hybridisation, reverse-transcription polymerase chain reaction (RT-PCR) was performed, to confirm the initial SAGE tag-to-gene transcript match and to provide a probe template for radio-labelling by PCR.

2.2.4.1 Reverse-transcription polymerase chain reaction (RT-PCR)

For RT-PCR, first strand cDNA synthesis was performed using Invitrogen's SuperScript II Reverse Transcriptase, plus components supplied with this enzyme. 1 μg total RNA was added to 1 μl oligo$(dT)_{12-18}$ (500 μg/ml) and dH_2O to a total volume of 12 μl. The mixture was heated to 70°C for 10 min then quickly chilled on ice. 4 μl 5X first strand buffer, 2 μl 0.1M DTT and 1 μl 10 mM dNTPs were added to the tube and incubated at 42°C for 2 min. 1 μl (200 U) SuperScript II RNase H^- Reverse Transcriptase was then added and incubated at 42°C for 50 min. The reaction was stopped by heating at 70°C for 15 min. In order to remove RNA complementary to the cDNA, 1 μl (2 U) *E coli* RNase H^- was added and incubated at 37°C for 20 min.

The cDNA was then diluted to a total volume of 100 μl and used as a template for amplification in PCR, as described in section 2.2.2.8. The starting volume for PCR was 1% of this first strand reaction, but up to 10% was used where amplification proved more difficult to attain. As before, complementary single-stranded oligonucleotide primers were designed using appropriate software (see Section 2.2.2.8 and Table 2.2). The presence and size of the expected PCR product was confirmed by agarose gel electrophoresis (see Section 2.2.2.6). If the DNA fragment appeared appropriate, it was then purified using a Qiagen column (Section 2.2.2.5) and sequenced using the original PCR primers individually (Section 2.2.2.10), to confirm the identity of the sequence. Thereafter the fragments could be used as templates for probes for Northern hybridisation (Section 2.2.4.2.2) or for cloning (Section 2.2.5.2).

2.2.4.2 Northern blotting

2.2.4.2.1 Preparation of own Northern blot

Northern blotting is a technique which identifies mRNA species that have been size-fractionated by gel electrophoresis. Northern blotting was performed using Ambion's NorthernMax™ Kit, according to manufacturer's instructions. A 1% denaturing agarose gel was prepared with the formaldehyde-containing Denaturing Gel Buffer and poured in a 15x15 cm casting tray to a thickness of 6 mm. 30 µg total RNA, in a 10 µl volume, was mixed with 3 volumes Formaldehyde Load Dye containing 10 µg/ml ethidium bromide.

The samples were from 19 gastro-intestinal tumour and normal tissues and cell lines, including the three SAGE samples (Sections 2.1.5.1 and 2.1.5.3, and Table 4.3). The total RNA was incubated at 65°C for 15 min, chilled quickly on ice and loaded on the gel, along with an RNA molecular weight marker. The gel was run in 1X MOPS Gel Running Buffer at 130V until the bromophenol blue dye front was close to the bottom of the gel. The gel was wrapped in plastic wrap, examined with UV light and photographed. The blotting stack of paper towels and 3MM blotting paper was overlaid with a wet BrightStar-Plus Membrane. The gel was positioned on top of the membrane then covered with a filter paper bridge extending into the Transfer Buffer reservoir and with the weighted gel casting tray. Downward transfer was for 90 min. The membrane was exposed to UV light to crosslink the RNA.

2.2.4.2.2 Preparation of DNA probe

DNA probes for the Northern blot were generated by linear amplification of the appropriate PCR product using Ambion's Strip-EZ™ PCR Probe Synthesis and Removal Kit. Ambion's Strip-EZ technology involves synthesizing a StripAble™ probe that is stable under the conditions used for hybridization and washing, but which is cleaved by a reagent in the Probe Degradation Buffer supplied with the kit. Degradation is specific to the probe, because cleavage occurs at a modified nucleotide that is incorporated during probe synthesis. The use of harsh stripping protocols is thus avoided and thus (in theory) the lifespan and consistency of the precious blots is increased.

The reaction components (all from the kit) were assembled on ice. 2 µl of each of the 10X PCR buffer, 10X dNTP solution and the antisense primer (10 µM stock) were mixed with 10 ng of the spin-column-purified PCR product. The volume was made up with dH_2O to 17 µl. Then 2 µl [α^{32}P]dATP and 1 U of *Taq* DNA Polymerase were added. The StripAble™ Probe was labelled by linear PCR amplification using a Thermo Hybaid Thermal Cycler. The cycling parameters were: preheated to 94.5°C; 30 cycles of 94.5°C for 30 s, 50°C for 1 min and 72°C for 1 min; then 4°C.

Probes were generated and used for: gastrin; the new gene CA11 (foveolin); prostate stem cell antigen (PSCA); lipocalin 2 (neutrophil gelatinase-associated lipocalin); intestinal trefoil factor (TFF3); prothymosin alpha; thymosin beta 10; and Id1 (inhibitor of differentiation 1). As a loading control, actin was used, generated from a purchased template.

2.2.4.2.3 Hybridisation of probe and visualisation

The ULTRAhyb™ prehybridisation/hybridization solution was preheated to 68°C. 10 ml ULTRAhyb was used per 100 cm^2 of membrane (25 ml for one 15x15 cm BrightStar Plus membrane). The Northern blot was prehybridized in a hybridisation tube in a roller oven at 42°C for at least 30 min. All of the radiolabelled probe was mixed with 1 ml of preheated ULTRAhyb™ then transferred immediately to the container with the prehybridized blot and mixed well. Hybridisation was performed overnight. After the incubation, the radioactive ULTRAhyb was disposed of appropriately.

The blot was then washed, using 20 ml wash solution per 100 cm^2 membrane for all washing steps, with agitation. The first two washes were with Low Stringency Wash Solution #1 (2X SSC, 0.1% SDS), each at RT for 5 min. The next two washes were with High Stringency Wash Solution #2 (0.1X SSC, 0.1% SDS), each at 42°C for 15 min. The blot was wrapped in plastic film to prevent drying. The blot was then visualized directly using a phosphoimager with the BioRad Gel Doc 1000 gel documentation system and Molecular Analyst software.

2.2.4.2.4 Stripping probe from the blot

For probes labeled using Ambion's Strip-EZ™ PCR Probe Synthesis kit, stripping was performed according to manufacturer's instructions. The blot was washed in 1X Probe Degradation Buffer at RT for 2 min then at 68°C for 10 min. The blot was then washed in 1X Blot Reconstitution Buffer at 68°C for 10 min. The blot was checked for residual probe by monitoring with a Geiger counter then stored wrapped in plastic film at -20°C.

2.2.4.3 Immunohistochemistry (IHC)

Immunohistochemistry was kindly performed by Iain Downie in the University Department of Pathology using the standard avidin-biotin complex technique. The formalin-fixed and paraffin-embedded tissue or cell sections were dewaxed in xylene and rehydrated through intermediate alcohols. The samples used were the same as for the Northern blot (see Section 2.1.5.1). Endogenous peroxidase was removed by immersing in 0.3% hydrogen peroxide for 10 min, then the slides were washed in water for 5 min. Endogenous biotin was blocked using avidin and biotin from the VECTASTAIN Universal Elite ABC-Peroxidase Kit kit.

The primary antibodies used were gastrin, lysozyme, MUC5AC, pS2, human spasmolytic polypeptide and cytokeratin 8. Where the supplier recommended antigen retrieval, this was performed with either: 0.1% trypsin / 0.1% calcium chloride at 37°C for 10 min; or pressure cooker microwaving in 1 mM EDTA pH 8.0, for 5 min at full pressure followed by 20 min cooling down. The sections were incubated in normal horse serum blocker for 15 min, followed by incubation with the primary antibody for 30 min at RT. Detection was by a universal rabbit/mouse biotinylated secondary and then by the avidin-biotin complex (VECTASTAIN Kit), according to instructions. Visualisation with 3,3-diaminobenzidine was enhanced in 0.5% copper sulphate. Negative controls were treated as per test sections but with normal horse serum instead of primary antibody.

2.2.5 Further investigation of new gastric gene

The new gastric gene, called CA11 or foveolin, was further characterised at the mRNA and DNA levels in human, mouse and other species. First *in silico* investigation on-line and using local software identified likely transcripts. At the mRNA level, the expression profile of the gene was evaluated in a range of tissues by Northern blotting and *in situ* hybridisation, using the cloned full-length cDNA fragment as a probe. The mRNA transcript in the human and mouse was structurally investigated by rapid amplification of cDNA ends (RACE). A human genomic clone was identified and used to map the gene to human chromosome 2. The genomic sequence for both the human and mouse was then obtained on-line. The DNA, mRNA and predicted proteins for the different species were analysed on-line and compared using local software.

2.2.5.1 Analysis of EST and mRNA/cDNA sequences for the new gastric gene

Sequences corresponding to the new gastric gene, foveolin, in the human, as well as in the mouse, rat and cow, were investigated in depth. Local analysis was performed using either the DNASTAR suite of programs, especially EditSeq and SeqMan, or Vector NTI software. On-line analysis involved mainly the NCBI websites, especially SAGEmap, UniGene and BLAST, plus the EBI websites (Section 2.1.13). The bioinformatics websites, procedures and programs used to investigate the new gastric gene are described in great detail in Chapter 5.1, in the context of the questions being addressed at each stage.

2.2.5.2 Cloning of RT-PCR product for use as a probe

TOPO TA Cloning® was kindly performed by Sharon Burns of O2. A 750 bp fragment corresponding to the full-length foveolin cDNA was ligated into the expression vector pCR®II-TOPO® vector using the TOPO TA Cloning® Kit, which provides a highly efficient, 5-min, one-step cloning strategy for the direct insertion of *Taq* polymerase-amplified PCR products into a plasmid vector. For chemically competent *E. coli*, 4 µl fresh PCR product was mixed with 1 µl salt solution and 1 µl pCR®II-TOPO® vector. The reaction was incubated at RT for 5 min then placed on ice. 2 µl of the cloning reaction was then used for bacterial transformation, according to manufacturer's instructions.

2.2.5.3 Plasmid DNA preparation

Plasmid DNA preparation was kindly performed by Sharon Burns of O2. The Qiagen plasmid midi/maxi kit and protocol were used. A single colony was picked from a plate and used to inoculate a starter culture of 2-5 ml LB medium containing selective antibiotic. The starter culture was incubated with shaking at 37°C for around 8 h. The starter culture was diluted into 500 ml selective LB medium and incubated with shaking at 37°C overnight. The bacterial cells were harvested by centrifuging at 6,000 rpm for 15 min at 4°C. The bacterial pellet was resuspended in 10 ml of Buffer P1. 10 ml of Buffer P2 was added, mixed by inversion, and incubated at RT for 5 min. 10 ml of chilled Buffer P3 was added, mixed by inversion, and incubated on ice for 15 min. The mixture was centrifuged at 20,000 g for 30 min at 4°C. The supernatant, which contains the plasmid DNA, was removed and re-centrifuged. The supernatant was again removed.

A QIAGEN-tip 500 column was prepared by applying 10 ml Buffer QBT then allowing the column to empty by gravity flow. The supernatant was applied to the QIAGEN-tip column and left to enter the resin by gravity flow. The column was washed with 2 x 30 ml Buffer QC. The plasmid DNA was eluted with 15 ml Buffer QF then precipitated by adding 10.5 ml room-temperature isopropanol, mixing and centrifuging immediately at 20,000 x g for 30 min at 4°C. The supernatant was discarded. The DNA pellet was washed with 5 ml room-temperature 70% ethanol and centrifuged at 20,000 x g for 10 min. The supernatant was again discarded. The DNA pellet was air-dried for 10 min then resuspended in TE buffer.

Thereafter the DNA plasmid preparation was restriction-enzyme digested with either *Kpn* I (antisense) or *Not* I (sense) to linearize the cDNA template and to avoid transcription of undesirable sequences. The template was then used to generate probes for Northern blotting, as before (Section 2.2.4.2.2), and for *in situ* hybridisation, as below (Section 2.2.5.4.1).

2.2.5.4 In situ hybridisation (ISH)

2.2.5.4.1 Preparation of riboprobe

Two types of digoxigenin-labelled riboprobe (RNA probe) were generated for use in *in situ* hybridisation (ISH). The first, the antisense riboprobe, is complementary to the target foveolin mRNA and therefore hybridises to it; this serves as the positive control and test. The second, the sense probe, is identical to the target foveolin mRNA and therefore should not hybridise; this serves as a negative control.

A full-length foveolin cDNA template was generated within the expression vector pCRII-TOPO and linearised as described above (Section 2.2.5.3). Antisense or sense digoxigenin-UTP-labelled riboprobes were then generated by reverse-transcribing the cDNA fragment using either T7 (antisense) or Sp6 (sense) RNA polymerase. Reverse-transcription was performed using the Roche DIG RNA Labeling Kit according to manufacturer's instructions. Briefly, 1 µl linearized template cDNA was mixed with 2 µl DIG/NTP mix containing RNA nucleotides and digoxigenin-11-uridine-5'-triphosphate (DIG-11-UTP), 2 µl 10X Transcription Buffer, 12 µl dH_2O, 1 µl RNase inhibitor and 2 µl RNA polymerase, and incubated at 37°C for 4 h.

2.2.5.4.2 Hybridisation of riboprobe and visualisation

In situ hybridisation was kindly performed by Rod Ferrier and Sharlene Butler in the University Department of Pathology.

A wide range of normal human tissues and their corresponding tumours, principally adenocarcinomas and their originating glandular epithelia, were examined as detailed previously in Sections 2.1.5.1, 2.1.5.2 and 2.2.1.3. Animal stomach was also examined (Section 2.2.1.3). Formalin-fixed and paraffin-embedded tissue or cell sections were dewaxed in xylene and rehydrated through alcohols to phosphate buffered saline (PBS). The slides were immersed in 0.2 N HCl then 0.3% Triton X 100 at RT for 15 min each on a shaker; after each step the slides were rinsed in PBS. The slides were incubated with

Proteinase K solution (100 μg/ml) in a moist box at 37°C for 30 min then rinsed in PBS. The slides were incubated with pre-hybridisation solution at 37°C for 1 hr, then the excess was poured off. 20-40 μl hybridisation buffer containing labelled antisense probe, at a 1:400 dilution, was dotted on coverslips and mounted on all slides except the negative control. The slides were placed in an Omnislide thermal cycler and allowed to hybridise overnight at 42°C.

On the second day, the slides were washed in 0.1X SSC at RT for 10 min, 0.1X SSC at 45°C for 30 min and 0.1X SSC at RT for 10 min. They were then rinsed in Dig. (1) Buffer at RT for 5 min. Anti-Digoxigenin-Alkaline Phosphatase conjugate in 1:2000 Dig. (1) Buffer and 10% Normal Swine Serum was applied to each section and incubated at RT for 2 h in a moist chamber. Excess antibody was washed off in Dig. (1) Buffer at RT for 10 min then rinsed in Dig. (3) Buffer at RT for 5 min. Slides were immersed in NBT/BCIP solution, with levamisole added at 2mM, overnight in a dark cupboard at RT.

On the third day, slides were washed in water and counterstained lightly with haematoxylin. The alternative counterstain, Alcian Blue, was used in a few cases to visualise acidic mucins in goblet cells in true or metaplastic intestinal epithelium. Coverslips were mounted on slides with melted glycergel and allowed to harden.

Normal stomach was used for both positive (antisense riboprobe) and negative controls. A positive reaction is indicated by blue-black staining (see Figure 5.5). During the initial validation of the new foveolin probe for ISH, three types of negative controls were used: hybridisation with sense (negative) foveolin riboprobe; hybridisation with antisense (positive) probes for renin and albumin, which are not normally expressed in the stomach; and simple omission of probe, with hybridisation buffer alone. In the main study, each run included positive and negative controls to confirm that any positive signal was indeed that of target foveolin mRNA.

2.2.5.5 Rapid amplification of cDNA ends (RACE)

Rapid amplification of cDNA ends (RACE) is a common method used to analyse, separately, the 5'- and 3'-ends of mRNAs. The SMART^{TM} RACE cDNA Amplification Kit was used to generate the 3' and 5' ends of the foveolin mRNA, according to the manufacturer's instructions. The oligonucleotide primers were designed using Vector NTI in accordance with advice given in the kit, and are shown in Table 2.3.

For synthesis of first-strand cDNA for RACE, two 0.5 ml microcentifuge tubes were set up, one for the 5' reaction and one for the 3' reaction. To both tubes, 1 μg (3 μl) total RNA was added. To the 5' tube was added 1 μl 5'-CDS primer and 1 μl SMART II A oligo. To the 3' tube was added 1 µl 3'-CDS primer A and 1 µl dH_2O. The reactions were incubated at 70°C for 2 min and cooled on ice for 2 min. The following was added to each reaction tube: 2 µl 5X First-Strand buffer, 1 µl DTT (20 mM), 1 µl dNTP Mix (10 mM) and 1 µl PowerScript Reverse Transcriptase. The contents were mixed and incubated at 42°C for 1.5 hr. The first-strand reaction product was diluted with 100 µl Tricine-EDTA Buffer and heated at 72°C for 7 min.

For the Rapid Amplification of cDNA Ends (RACE) reaction, a PCR Master Mix was prepared containing the following reagents per 50 µl reaction: 34.5 µl dH_2O, 5 µl 10X Advantage 2 PCR Buffer, 1 µl dNTP Mix (10 mM) and 1 µl 50X Advantage 2 Polymerase Mix, to give a total volume of 41.5 µl. This was mixed and distributed into the different PCR reaction tubes. To the 5'-RACE reaction was then added 2.5 µl of the 5'-RACE-Ready cDNA, 5 µl Universal Primer Mix (UPM) (10X) and 1 µl Gene Specific Primer 1 (GSP1) (10 µM). To the 3'-RACE reaction was then added 2.5 µl of the 3'-RACE-Ready cDNA, 5 µl Universal Primer Mix (UPM) (10X) and 1 µl Gene Specific Primer 2 (GSP2) (10 µM). The protocol recommended a number of control RACE reactions: negative controls included the use of the Universal Primer Mix without Gene Specific Primers, and of the "wrong" Gene Specific Primer, while positive controls included the use of both Gene Specific Primers in one reaction and of supplied primers specific for the ubiquitously expressed TFR (transferrin receptor). Each PCR reaction was overlaid with 2 drops of mineral oil then cycled with Touchdown PCR, as follows: 5 cycles of 94°C for 30 s and 72°C for 3 min; 5 cycles of 94°C for 30 s, 70°C for 30 s and 72°C for 3 min; and 25 cycles of 94°C for 30 s, 68°C for 30 s and 72°C for 3 min.

Both 5'- and 3'-RACE reactions should produce a single clear product band. The PCR products were then purified using Qiagen columns (Section 2.2.2.5) and TOPO TA® cloned (Section 2.2.5.2), again by Sharon Burns. Five different colonies were selected, plasmid DNA prepared (Section 2.2.5.3) and sequencing performed (Section 2.2.2.10). The sequence files were analysed and assembled using SeqMan.

2.2.5.6 Probing of genomic library

Identification of a clone containing the human genomic sequence for the new gastric gene was performed using the human PAC library RPCI1. This is a genomic library within PACs (plasmid artificial chromosomes) was obtained spotted on seven high-density gridded filters on 22 x 22 cm nylon membranes. The source was a normal male blood donor, and the insert size is about 110 kb. Each clone has been spotted twice to give 36,864 (18,432 x 2) spots on each membrane, with 7 filters covering the whole library.

The previously described radiolabelled full-length human cDNA probe (Section 2.2.4.2.2) was used to probe the PAC libraries. The membranes were placed in rotisserie bottles, with two or three membranes, separated by nylon mesh, per bottle. Prehybridisation, hybridisation and washing were performed using ULTRAhyb™ solution as for the previous Northern blots (Section 2.2.4.2.3) except that where heat was used, the temperature applied was not 42°C but instead 65°C. After washing, the filters were wrapped in plastic wrap, placed in film cassettes and exposed to X-ray film for between 2 to 72 hours (usually overnight was best) at -70°C. The position of the positive clones was interpreted following the protocol and orientation guide supplied.

A single positive clone was found and identified as number 291-N4. This clone was obtained and a DNA plasmid preparation was performed, according to supplier's instructions, but basically using similar methods as before (Section 2.2.5.3). The cloned PAC DNA was then used as a probe for fluorescent *in situ* hybridisation (FISH) and was also used for genomic sequencing.

2.2.5.7 Fluorescent in situ hybridisation (FISH)

Fluorescent *in situ* hybridisation (FISH) for the new gastric gene, foveolin, was kindly performed by Sharon Burns in O2 group. The foveolin probe for FISH was labelled using the Vysis Nick Translation Kit and SpectrumGreen dUTP. 1 μg of plasmid DNA (Section 2.2.5.6) was mixed with dH_2O to a total volume of 17.5 μl. To this was added: 2.5 μl 0.2 mM SpectrumGreen dUTP; 5 μl 0.1 mM dTTP; 3 μl of each of 0.3 mM dATP, dCTP and dGTP; 5 μl 10X Nick Translation Buffer; and 10 μl Nick Translation Enzyme. The reaction was incubated overnight at 16°C, then stopped by heating at 70°C for 10 min and chilled on ice. The labelled probe was ethanol precipitated and resuspended in 6 μl 50% formamide hybridisation mix (2X SSC, 500 μg/ml salmon sperm DNA, 10% dextran sulphate and 50% formamide).

On the seond day, the chromosome preparation was dropped on to a glass slide from a height then fixed for 1 h in methanol / acetic acid (3:1) at RT and air-dried. The slide was incubated for 1 h in 100 μg/ml RNase in 2X SSC at 37°C, then rinsed in 2X SSC. The chromosome preparation was then digested in pepsin (0.01% in 10 mM HCl) solution for 10 min at 37°C and rinsed in water. The slide was fixed for 10 min in Streck Tissue Fixative at RT. The slide was then dehydrated in 2 x 2 min 70% ethanol, 2 x 2 min 100% ethanol, and air-dried. 30 μl 70% formamide in 2X SSC was added to the slide which was then covered with a cover slip and denatured at 80°C for 2 min on the Omnislide Thermal Cycler. At the same time, the labelled probe and the chromosome 2 paint (6 μl of each) were also denatured, at 75°C for 5 min. The slide was again dehydrated. The denatured probe and paint were applied to the slide, covered with a coverslip, sealed with cow gum, and hybridised at 37°C overnight.

On the third day, the cow gum and coverslip was removed by soaking in 2X SSC. The chromosome preparations were washed in 0.4X SSC / 0.3% NP-40 at 75°C, with agitation for 2 s then standing for 2 min. The slides were then washed in 2X SSC / 0.1% NP-40 at RT, with agitation for 2 s, then standing for 1 min. The slides were dehydrated in 2 x 2 min 70% ethanol, 2 x 2 min 100% ethanol, and air-dried. The slides were mounted in Vectashield with 0.3 μg/ml propidium iodide (PI) / 0.1 μg/ml 4',6-diamidino-2-phenylindole (DAPI), and sealed with nail varnish. The stained chromosome preparations were examined under a fluorescence microscope and photographed.

2.2.5.8 Identification and analysis of genomic and protein sequences

(See also previous Section 2.2.5.1.) The next aim was further characterisation of the genomic sequence of the new gastric gene in human and mouse. The human PAC genomic clone was sequenced (Section 2.2.2.10) using the primers listed in Table 2.3 (Section 2.1.8.2). The resulting genomic sequences were assembled using DNASTAR's SeqMan, as described in detail later in Section 5.1.2.3, but the clone turned out to contain only part of the gene.

The full human genomic clone was thus identifed and down-loaded from the Celera web-site. The mouse genomic sequence was identified and down-loaded piece-meal from the mouse Whole Genome Shotgun sequencing reads in the Trace Archive at Ensembl. Comparisons were again performed using BLAST and MultAlign; or by hand (!). This is explained in more detail later in Sections 5.1.2.4, 5.1.2.5 and 5.1.2.6. Thanks are due to Steve Bryce for his kind advice on these genomic searches.

The protein sequences for the human, mouse and cow were predicted using EditSeq, compared with ClustalW and analysed using the Network Protein Sequence Analysis programs and PSORT (Section 2.1.13.1), as explained in more detail later in Section 5.1.3. Searches for protein homology and motifs were performed using BLAST and the various protein databases and tools available at the website of the European Bioinformatics Institute (EBI).

Table 2.1: Clinico-pathological details of gastric adenocarcinoma samples

This table provides clinico-pathological information on the gastric adenocarcinomas used in this research (see Section 2.1.5.1). Tumours number 1 and 2 were used for SAGE, as was a sample of normal (non-neoplastic) gastric antral mucosa taken from a Whipple's resection from a patient with pancreatic (but not gastric) adenocarcinoma. For SAGE, RT-PCR and Northern blotting, RNA was prepared from the frozen samples. For immunohistochemistry and *in situ* hybridisation, the formalin-fixed, paraffin-embedded tissue was used.

Background information on the significance of the tumour site, anatomical and histological subtypes, tumour grade (histological differentiation) and tumour stage is provided in the Introduction, Section 1.4. Note that here, T and N refer to classical tumour staging (local Tumour spread and presence of lymph Node metastases), as described in Table 1.1. However, throughout the rest of the paper, T and N indicate the SAGE Tumour and Normal samples.

Tumour number	Tumour site	Anatomical subtype	Histological subtype	Histological differentiation	Tumour stage
1	Body	Distal	Intestinal	Well	T1 N0
2	Antrum	Distal	Intestinal	Moderate	T3 N0
3	Antrum	Distal	Intestinal	Moderate	T3 N1
4	Antrum and body	Distal	Diffuse	Poor	T1 N0
5	Antrum	Distal	Diffuse	Poor	T3 N2
6	Gastro-esophageal junction (GEJ)	Proximal	Intestinal and solid	Poor	T3 N2
7	GEJ	Proximal	Diffuse and intestinal	Poor	T3 N1
8	GEJ	Proximal	Intestinal, solid and basaloid	Poor	T2 N0

Table 2.2: Oligonucleotide primers used in RT-PCR for SAGE validation

This table lists the oligonucleotide primers used to investigate the new gastric gene, foveolin, as described in Section 2.2.5 and Chapter 5.

The primers named human cDNA were designed from the sequence generated by assembling the down-loaded ESTs. They were used in various combinations for initial amplification and characterisation of the human cDNA sequence, by reverse-transcription polymerase chain reaction (RT-PCR) and rapid amplification of cDNA ends (RACE).

To validate and expand the SAGE profiles, selected transcripts were studied in a wider panel of 19 gastro-intestinal tumour and normal tissues and cell lines. Where the genes had been minimally characterised in the stomach, the method of choice was Northern blotting for mRNA. Prior to Northern hybridisation, reverse-transcription polymerase chain reaction (rtPCR) was performed, to confirm the initial transcript match and to provide a probe template for radio-labelling by PCR.

Name	Oligonucleotide primer sequence
gastrin sense	GCCCAGCCTCTCATCATC
gastrin as	GGGGACAGGGCTGAAGTG
TFF3 sense	CAGTCCTGAGCTGCGTCCCG
TFF3 as	CAGGCACGAAGAACTGTCCTCG
thymosin beta10 longer sense	GTGGGAGCACCAGGATCTC
thymosin beta10 longer as	GAATTTGGCAGTCCGATTG
prothymosin alpha upper	ACACCAGCTCCGAAATCACC
prothymosin alpha lower	TCATCCTCGTCGGTCTTCTG
prostate stem cell antigen sense	CTGCTTGCCCTGTTGATGGC
prostate stem cell antigen as	TGCGTTAGGATGTGCCTCAGG
Id1 longer sense	GGACGAGCAGCAGGTAAACG
Id1 longer as	CACACGAGTGGAATCCCACC
hsp90a sense	ACCCAGACCCAAGACCAACCG
hsp90a as	ATTTGAAATGAGCTCTCTCAG
NGAL (lipocalin 2) sense	GGACTCCACCTCAGACCTGATC
NGAL (lipocalin 2) as	CGATACACTGGTCGATTGGGAC

Table 2.3: Oligonucleotide primers used in investigation of new gastric gene

This table lists the oligonucleotide primers used to investigate the new gastric gene, foveolin, as described in Section 2.2.5 and Chapter 5.

The primers named human cDNA were designed from the sequence generated by assembling the down-loaded ESTs. They were used in various combinations for initial amplification and characterisation of the human cDNA sequence, by reverse-transcription polymerase chain reaction (RT-PCR) and rapid amplification of cDNA ends (RACE).

The human cDNA primers were also used to "walk" into the introns during characterisation of the human genomic sequence, for which the primers named genomic were also used. Many of the latter, named upto307, turned out unfortunately to be located within PAC vector sequence, as described in Chapter 5.

The primers named mouse cDNA were also designed from the sequence generated by assembling the down-loaded ESTs. They were used in various combinations for initial amplification and characterisation of the mouse cDNA sequence, by reverse-transcription polymerase chain reaction (RT-PCR) and rapid amplification of cDNA ends (RACE).

Name	Oligonucleotide primer sequence
human gastric specific cDNA	
human gastric specific upper	TCCATTCAATCCCTTGATGC
human gastric specific lower	CCGAACTTGCTCAGGTCATC
human gastric specific 001 sense	CCTCTGTCCACTGCATTTC
human gastric specific 462 sense	GGATTCCAACATACATGGC
human gastric specific 532 sense	GACCAGTGTACTATGGATTGTGG
human gastric specific 620 sense	CATCTGAATATGCTGTGCAGA
human gastric specific 109 as	CCAGCATTGTTGTTGTCATCATTG
human gastric specific 237 as	GAGTCTGGTTGCAGCAAAGC
human gastric specific 307 as	GGGCATGACTTCCTTGTTC
human gastric specific 736 as	GCTAAATGATTTTATTGAAACTTAA
human gastric specific RACE upper	CCACCTCCCAAGGGCCTGATGTACTCAG
human gastric specific RACE lower	CCTCAGCCATGTATGTTGGAATCCCACG
human gastric specific genomic only	
human gastric specific genomic 3'1 as	GCCTTGATGACCATCCCTGC
human gastric specific genomic mid1 as	GGGAATGACTGTGATCTTCTCG
human gastric specific contig 462-736 as2	GTGACCTCTCTGTACCTCAGTTTCC
human gastric specific contig low-high as2	GCGGTTGTAGAGGTAGGTGTG
human gastric specific genomic upto307 as1	GACATTTGTGAGGACAGACTGC
human gastric specific genomic upto307 as2	GCTGATAATTCCAGAGTTGGGTC
human gastric specific genomic upto307 as3	CAGCATTATTGAGGGTATCTGC
human gastric specific genomic upto307 as4	TACACGCCATGATATGCTGC
human gastric specific genomic upto307 as5	CTCTCTTACCAATTAATATTGTC
human gastric specific genomic upto307 as6	CATTTACATTTCCCGAGTTTTAGG
human gastric specific genomic upto307 as7	ACATTTCCCGAGTTTTAGGTG
human gastric specific genomic upto307 as8	GGTCGTAAAACCTCCTTTGAAAAC
human gastric specific genomic upto307 as9	GGCAAGGCAAAAGGATGTTC
human gastric specific genomic upto307 as10	CCACAATCCCTTGAACATAGTG
human gastric specific genomic upto307 as11	TAGTCAATTCGGGAGGATCG
human gastric specific genomic upto307 as12	GAGGAGCGACTCAAGCCTTC
human gastric specific genomic upto307 as13	TACACGCCATGATATGCTGC
human gastric specific genomic upto307 as14	CCCTAAACAGCAGCAAAACG
human gastric specific genomic upto307 as15	GCAAAACGCTGACGGAACAG
human gastric specific genomic upto307 as16	TGTTTGCTCTTGCGGGAAGC
human gastric specific genomic upto307 as17	TGAAAGATCAGACGCAAGAATG
human gastric specific genomic upto307 as18	CGCAGGTAAATGTGTCAAAATC
human gastric specific genomic upto307 as19	CGGAGTGGAAGATCACAAAAC
human gastric specific genomic upto307 as20	CGGAGACGGAAAAACATATCAG
human gastric specific genomic upto307 as21	GATGAAGGCAATTATACATCCG
human gastric specific genomic upto307 as22	CCATACGCTGAGAGACCCTC
human gastric specific genomic upto307 as23	GCCATAAATACCTTGGATTCG
human gastric specific genomic upto307 as24	GGATACCAAGGCGAAGAATC
mouse gastric specific cDNA	
mouse gastric specific 20 sense	CTGATCCTCTGCTCCACCAC
mouse gastric specific 710 as	GTGGGTTTATTGAGCCTTAAAGAAG
mouse gastric specific RACE sense	CGGAAATGGACAGCATTCGGTGAGCATC
mouse gastric specific RACE as	CAGGTCCTCCACTCTGGTAGGGTTGACGG

3 Setting up SAGE and its optimisation

3.1 Results

Before proceeding to the results of the gene expression profiling in the stomach, it is worth considering the method used in more detail. Serial analysis of gene expression (SAGE) is a large-scale mRNA expression profiling technology with immense potential for producing valuable data-sets. When this project began, however, SAGE was a new technique, developed in the USA. I was the first researcher to use SAGE locally. Its translation to our laboratory presented challenges: its set-up took time, effort and (very gradually acquired) expertise, mainly for the laboratory work but also, to a lesser extent, for the bioinformatics.

In this chapter, first the principles underlying SAGE are presented. Then the issues involved in the set up and optimisation of SAGE are discussed. During the course of the project, improved protocols were published in the literature and developed locally, and their integration into this project is also described. The chapter thus incorporates both my personal experience of SAGE and on-going developments in the field.

3.1.1 Principles underlying Serial Analysis of Gene Expression (SAGE)

Serial analysis of gene expression (SAGE) produces comprehensive, quantitative and reproducible gene expression profiles. It was originally described in 1995 by Velculescu *et al* from the Kinzler and Vogelstein laboratory in Johns Hopkins University (Velculescu *et al.* 1995; Zhang, L *et al.* 1997).

SAGE is based on generating clones of concatenated (linked) short sequence tags derived from mRNA from the target cells or tissue. Each tag is 9 or 10 bp long and represents a single mRNA. This short nucleotide sequence should, in theory, contain sufficient information to identify the transcript uniquely, providing that the tag is taken from a defined position within the transcript. For example, the 9 bp tags which were used in this project should be able to distinguish 262,144 (4^9 = 4 possible nucleotides at 9 possible positions) transcripts, assuming that the distribution of nucleotides at the tag site is random. In theory, each clone insert could contain up to 40 tags joined serially, although in practice the inserts which I generated were usually significantly shorter (see Section 3.1.2.4). Sequencing of multiple concatenates therefore describes the pattern and abundance of mRNA, with an improvement in efficiency of up to 40-fold compared with conventional analysis of expressed sequence tags (ESTs), in which each clone insert contains only a single mRNA sequence. The boundaries between the tags are recognised by a 4 bp nucleotide sequence (CATG) corresponding to the recognition site of the *Nla*III restriction enzyme used during SAGE. The mRNA transcript corresponding to the short SAGE tag is identified from the Genbank genetic databases using appropriate software.

Figure 3.1 provides a schematic diagram of how SAGE is put into practice. Figure 3.2, Figure 3.3, Figure 3.4 and Figure 3.5 show representative examples of the many gel electrophoresis steps used in SAGE. The theory underlying SAGE, starting from the input mRNA through to obtaining the ditags, is as follows. Double stranded cDNA is synthesized from mRNA using a biotinylated oligo(dT) primer. The cDNA is then cleaved with the so-called Anchoring Enzyme, in this case *Nla*III, which is a restriction endonuclease with a 4 bp recognition and cleavage site (CATG). On average *Nla*III would be expected to cleave every 256 bp (4^4 = 4 possible nucleotides at 4 possible positions, assuming that the distribution of nucleotides is random). Most mRNA transcripts are considerably longer than 256 bp: their average length is 2.5 kb (Strachan *et al.* 1999) but range from a few hundred to many thousands of bases (Sambrook *et al.* 1987). *Nla*III should therefore cleave most transcripts at least once. The 3' end of the cleaved biotinylated cDNA is then captured by binding to streptavidin-coated magnetic beads. This process creates a unique site on each mRNA transcript that corresponds to the *Nla*III restriction site which is the most 3', that is, located closest to the polyA tail.

The cDNA is then divided in half and ligated via the anchoring restriction site to one of two oligonucleotide linkers containing a recognition site for *BsmF*I, the so-called Tagging Enzyme, which is a type IIS restriction enzyme. The latter cleave at a defined distance, up to 20 bp away from, their asymmetric recognition sites, leaving a sticky-end. The oligonucleotide linkers are designed so that cleavage of the ligation products with *BsmF*I releases the linker plus a short piece of the original cDNA, which is the SAGE tag. The released linker-tag combination is first blunt-ended at the residual sticky-end created with *BsmF*I. The two pools of linker-tags are then ligated to each other and amplified by PCR with linker-specific primers (Figure 3.3). As well as amplifying the tag sequences for further manipulation, this step serves two other functions. First, it enables orientation and punctuation of the tag sequences. The amplification products contain two tags (one ditag) linked tail to tail (3' end to 3' end), flanked by CATG, the recognition site of the Anchoring Enzyme *Nla*III. Second, analysis of the ditags, which are formed before amplification, allows any PCR-induced bias to be avoided: repeated ditags potentially produced by biased PCR can be identified after sequencing and excluded from the final analysis.

Cleavage of the PCR product with *Nla*III releases the sticky-ended 22-26 bp ditags which are then isolated and concatenated (linked) by ligation (Figure 3.4). The concatemerised ditags are cloned (Figure 3.5) and sequenced (Figure 3.6). The SAGE sequences are then analysed using the SAGE program (Figure 3.7). Interesting tags can be further investigated using NCBI's on-line bioinformatics tools (Figure 3.98 and 3.9). A detailed explanation of the SAGE software and on-line bioinformatics tools used is provided in Sections 2.2.3.2 and 3.1.3 and in the legends accompanying Figure 3.7, Figure 3.8 and Figure 3.9.

3.1.2 The laboratory work in SAGE was difficult and time-consuming

The laboratory work in SAGE proved difficult and took many months to establish. This problem was not limited to our laboratory and indeed there are many centres worldwide which have attempted but never succeeded with SAGE and which abandoned it in favour of cDNA or oligonucleotide microarrays. This is partly because, although the steps in SAGE are individually simple, when they are linked together, there are many possible sources of error. Unfortunately, there are only two main stages at which the success of the SAGE procedure can be easily ascertained: obtaining the 102 bp PCR product and then obtaining appropriately sized PCR products of the cloned concatemers. The various technical problems encountered and, where identified, their solutions, are discussed in the following sections.

3.1.2.1 The quantity and quality of input mRNA was crucial

Initially I used locally obtained clinical tissues and, even after a number of months, could not achieve the 102 bp PCR product. As far as was possible, I checked whether each stage of SAGE prior to the PCR step was working as expected. Where necessary, alternative methods were tried and, in some cases, substituted. Most of the individual steps appeared to be functional, but through these checks, general discussions in the laboratory, and consultation of the academic and commercial literature, it became obvious that degradation of the input RNA could well be a problem.

In order to get SAGE working in the first place, I therefore purchased mRNA from Clontech, derived from normal whole human stomach. This mRNA eventually yielded both the 102 bp PCR product (Figure 3.3) and then the sequenced cloned concatemers (Figure 3.6), six months after starting. This obviously gave me confidence that I could perform SAGE. My original hope had been that this sample could be used to represent "normal stomach", but unfortunately further enquiry revealed that the Clontech mRNA was derived from samples containing the entire wall of the stomach, including mucosa, fatty submucosa, muscle and peritoneal surface, not just the mucosa, which is what defines the different areas of the gastro-intestinal tract (see Section 1.2.1) and from which carcinomas are derived (see Section 1.4.4). Clontech were also unable to specify from which area of the stomach the sample was taken, whereas my intention was to use mucosa from the distal stomach, the antrum, as the normal comparison for the distal tumours.

I therefore returned to my attempts to isolate and use mRNA from the local clinical samples for SAGE. It took a further eight months to start to obtain good quality clones from Tumour 1. Tumour 2 took only another two months but the normal sample was more tricky and took a further eight months. (Obviously the large-scale sequencing of the clones took much longer.)

The original SAGE protocol had recommended isolating total RNA with Qiagen's RNeasy® Midi Kit then purifying the mRNA with Invitrogen's MESSAGEMAKER™ Kit. Other strategies were also attempted: Ambion's Poly(A)Pure™ Kit was used to capture mRNA directly from powdered tissue lysates; Dynal's Dynabeads® Oligo(dT)$_{25}$ (oligo(dT)-coated magnetic beads) were used to purify mRNA both from total RNA and directly from tissue lysates; and Miltenyi Biotec's magnetic mRNA Isolation Kit (also containing oligo(dT)-coated magnetic beads) was applied to total RNA. None of these methods was successful in leading to a 102 bp PCR product.

However, most of these initial purification attempts had been undertaken without running analytical gels of the RNA from the clinical samples, partly to minimise sample loss and partly because, with some methods, the sample was taken from powdered tissue directly to mRNA without any intervening total RNA preparation. When the integrity of the input RNA was subsequently called into question, I proceeded to prepare total RNA from these tissues and check it by agarose gel electrophoresis. It became apparent that the RNA in a significant number of samples was in fact badly degraded, with smears, often weak, instead of strong, distinct ribosomal RNA bands (see Figure 3.2).

In retrospect, my initial omission of the analytical gels had obviously been a mistake, and would in any case have used only a tiny fraction of any total RNA sample. When I later extracted total RNA from cultured cell lines, gel electrophoresis always yielded clean ribosomal bands. The problem with the clinical tissues was therefore assumed to be due not to contamination with exogenous RNases, for example from my skin, which would be expected to affect all samples equally, but to endogenous RNases, in the gastric tissues themselves, which were presumably activated during and after surgery, before the sample could be dissected and snap-frozen. Their presence in abundance in the stomach, as in the pancreas (Sambrook *et al.* 1987), would not be unexpected given its normal digestive function and hence high enzyme content.

Thereafter every RNA sample was prepared in stages via total RNA so that its integrity could be checked by gel electrophoresis. TRIZOL® Reagent was used successfully (Section 2.2.2.2) and careful RNase-free techniques were applied at all times (Section 2.2.2.1).

3.1.2.2 Alternative components were tried in the pre-PCR steps of SAGE

For cDNA synthesis (Section 2.2.3.1.2), the original Johns Hopkins' protocol used Invitrogen's cDNA Synthesis Kit containing M-MLV reverse transcriptase. This enzyme has now been superceded by the same manufacturer's Superscript II RNase H^{-} Reverse Transcriptase, which lacks RNase H activity, preventing degradation of RNA molecules during first-strand cDNA synthesis, but which retains full polymerase activity. It is thus said to be four-fold more efficient, is now recommended by the Johns Hopkins workers, and was used in my later experiments.

The restriction enzyme *Nla*III is critical to SAGE where it is used in two separate steps to create the 22-26 bp ditags (Sections 2.2.3.1.3 and 2.2.3.1.11). Its activity, however, sometimes seemed to be low, to judge by the amount of enzyme and length of time required for complete digestion of the 102 bp ditags. This problem could be overcome by using more *Nla*III for a longer time. Possible causes could include batch-to-batch variability and incorrect storage at temperatures above –70°C, although I took care to avoid the latter. Impurities in the DNA preparation have also been shown to inhibit *Nla*III activity, and can be removed through the use of Qiagen spin columns (Angelastro *et al.* 2000).

For binding of the biotinylated cDNA to a streptavidin-coated surface, the original SAGE protocol recommended Dynal's Dynabeads® M-280 Streptavidin (Section 2.2.3.1.4). Promega's Streptavidin MagneSphere® Paramagnetic Particles were also tried but did not result in any improvement.

For blunt-ending (polishing) the sticky ends of the SAGE tags (single linker plus tag) created using *BsmF*I, the original SAGE protocol recommended T4 DNA Polymerase (Section 2.2.3.1.7). I had already been trying other DNA polymerases, including *Pfu* DNA Polymerase, when a Johns Hopkins' protocol update emerged with its preference for DNA Polymerase I Large Fragment (Klenow), which worked well.

3.1.2.3 PCR amplification of the 102 bp ditags was difficult

Initially, PCR amplification of the 102 bp ditags was difficult to achieve and not robust, and thereafter it was very prone to cross-contamination (Section 2.2.3.1.9).

The Johns Hopkins' protocol describes a complex and slightly unusual PCR reaction mixture, with high concentrations of $MgCl_2$, nucleotides and primers (its final reaction mix contained 6.7 mM $MgCl_2$, 1.5 mM nucleotides and 7 nM primers) and also including DMSO. It proved difficult to get the PCR working in the first place and then to optimise and maintain its performance. My previous experiences had suggested that the Qiagen kits worked well and their HotStarTaq kit yielded PCR product more reliably and in greater quantities than the Johns Hopkins recipe (its final, very different, reaction mix contained 2.75 mM $MgCl_2$, 0.2 mM nucleotides and 1.75 nM primers) .

Once the 102 bp ditag PCR product was obtained, PCR cross-contamination became a major problem, with undesirable ditag products appearing in the ligation and/or PCR negative control reactions. Its elimination took months and any SAGE libraries created in the interim had to be discarded since the PCR products thus obtained had not verifiably originated from the appropriate tissue sample. Cross-contamination was avoided by the use of: the Qiagen kit, with reactants newly opened each time; pre-PCR-only pipettes, pipette tips etc; and liberal use of Ambion's DNAZap solution and/or ultraviolet light to decontaminate benching and equipment.

3.1.2.4 It proved difficult to obtain long concatemer inserts in most clones

Once the 102 bp ditag PCR product is obtained, a large-scale reaction is performed to obtain sufficient quantities from which the 22-26 bp ditags can then be obtained by digestion with *Nla*III. Thereafter the short ditags are isolated from the linker oligonucleotides and residual PCR products. The 22-26 bp ditags are ligated with the aim of producing long (>500 bp) concatemers which must then be efficiently separated out, cloned and sequenced. Purification of the ditags and of the concatemers involves, once again, polyacrylamide gel electrophoresis, examples of which are shown in Figure 3.4. Many of these steps caused problems, with the common end result being only a small number of clones containing only short concatemer inserts.

3.1.2.4.1 Problems with the generation of long concatemers

Sometimes the problem was simply that long concatemers were not being generated, with most molecules remaining as single 22-26 bp ditags or forming only small multiples thereof. This usually related to the quantity and purity of the ditags.

Part of the solution involved adjustment of the ditag ligation conditions with the aim of promoting the forward reaction. The various adjustments included: decreasing the reaction volume; adding polyethylene glycol (PEG) (which effectively also decreases the reaction volume); allowing the reaction to proceed for longer; and trying ligase enzymes from different suppliers and using them at higher concentrations.

Another part of the solution involved increasing the quantity and purity of the input ditags. The first Johns Hopkins' protocol described only a single gel purification step: the pooled 102 bp ditag PCR products were directly *Nla*III digested before being run out on the gel to give the 26 bp ditags for isolation, plus remaining linkers. A further purification stage involves running out the pooled 102 bp ditag PCR products for isolation prior to *Nla* III digestion. I had already added this step by the time the modification appeared in the next Johns Hopkins' protocol.

In addition, contamination of the ditags with residual linker oligonucleotides may "poison" the ligation reaction, preventing further extension and also cloning of the concatemer chain. A novel approach to this problem involves the use of biotinylated primers for PCR of the 102 bp ditags (Powell 1998). After *Nla*III digestion, the unwanted linkers and undigested PCR products bear biotin and can be removed by washing with streptavidin-coated beads (Powell 1998). This streptavidin-biotin purification step is usually performed in addition to the two gel purification steps. Although other researchers, including the current Johns Hopkins' protocol, recommend this step, perhaps unexpectedly in my hands it did not effect any obvious improvement.

3.1.2.4.2 Problems with the separation out of long concatemers

In some experiments, it seemed that long concatemers were being generated but that their separation from the smaller species was poor, such that the eventual clone inserts mostly originated from the smaller (and therefore more easily cloned) fragments. This requires better separation by molecular weight. One modification is to use the same 8% PAGE separation as described in the original Johns Hopkins' protocol but to heat the concatemer sample prior to its loading to denature and separate the smaller species (Kenzelmann *et al.* 1999). In my hands, this did appear to provide slightly better results.

Others have used agarose gels, as described as an alternate procedure in the most recent Johns Hopkins protocol. Still other options include the use of various column formats, such as Amersham's MicroSpin™ columns, containing Sephacryl™ HR resin (Sephacryl S-400 HR) for spin-column chromatography, or QIAquick® columns. Having tried all these, I found that the QIAquick® columns gave results as good as 8% PAGE, in a much shorter time and with less risk of sample loss.

3.1.2.4.3 Problems with abundant contaminating clones lacking inserts

A further problem was that many of the selected resulting clones appeared to lack inserts. The Johns Hopkins' protocol describes the cloning of concatemers using Invitrogen's Zero Background™ Cloning Kit. The pZErO®-1 vector in this kit contains a lethal gene which is disrupted by DNA insertion, so that in theory only positive recombinants should grow, with bacterial selection using the antibiotic Zeocin. However, I found that although most colonies did contain vector, judging by the success of PCR amplification, many lacked a sizeable insert. The lack of bacteria without vector suggested that antibiotic selection with Zeocin was indeed working, although this would have been one potential source of problems, since this antibiotic is labile and its activity is sensitive to light, pH, salt concentration and temperature. Sequencing of the vector-only clones confirmed the lack of even a small insert (such as a single 22-26 bp ditag). The reason for the survival of vector-only clones was therefore unclear, although nuclease contamination resulting in damage to, and inactivation of, the lethal gene would be one explanation. Discussions with Invitrogen shed no further light on the problem.

I therefore tried a second cloning vector, pGEM®-3Zf(+), from the ABI PRISM® BigDye™ Terminator Cycle Sequencing Ready Reaction Kit. Unlike pZErO®-1, bacteria containing this vector are selected for by ampicillin-resistance and positive recombinants are identified through blue/white colony selection following IPTG/X-gal induction. PCR was performed with the same M13 forward and reverse primers as previously. Unfortunately, the original problem persisted: many white (i.e. selected) colonies lacked inserts of appropriate size.

Because the problem often seemed to be re-circularised vector lacking insert, treatment with calf intestinal alkaline phosphatase was also tried. This removes the 5'-phosphate from linearised vector and thus should inhibit vector self-ligation and re-circularisation without insert. However, in my hands, this treatment was ineffective.

I was therefore still faced with the problem of relatively inefficient SAGE concatemer cloning. From each 96-well plate of colony PCRs, thirty (if everything had worked well) and less than ten (if not, which was unfortunately much more common) colonies contained vector inserts of a size worth sequencing. This persisted despite months of attempted optimisation. Eventually I had to decide whether to continue to attempt to achieve better cloning (and in doing so not proceed with sequencing) or simply to accept that the cloning was relatively inefficient and to proceed with it as was and obtain some usable insert sequences. I chose the latter approach.

3.1.2.5 Large-scale PCR and sequencing of cloned concatemer inserts

The original Johns Hopkins protocol described colony PCR reactions in single microcentrifuge tubes, followed by manual precipitations with isopropanol, to remove residual primers, then sequencing with one of the primers previously used for the PCR. Such individual reactions were time-consuming, so I changed to a large-scale format. For the colony PCR reactions, 96-well plates were used. For the DNA purification, PCR with Qiagen's QIAquick 8 PCR Purification Kit (strips of 8 spin-type columns) was used with their QIAvac 6S (an 8 by 6 column gridded vacuum purification system). This purification method produced PCR product of which much less was needed for subsequent sequencing than with the isopropanol precipitation; larger amounts gave rise to poorer quality sequence, which was occasionally heterogeneous, presumably due to the contaminating presence of both primers, or possibly of more than one insert species.

3.1.3 Bioinformatics

By comparison with the difficulties encountered in the "wet-lab" part of SAGE, the problems in the bioinformatics were minor.

After sequencing the colony PCRs, the SAGE concatemer sequence files are analysed using the SAGE program in combination with NCBI's Genbank databases and Microsoft Access and Excel database programs. This procedure is explained in detail in Figure 3.7 and its accompanying legend. Briefly, the creation of a SAGE file first requires specification of the project parameters: in these experiments, the anchoring enzyme is *Nla*III, with its CATG recognition site; and the tag length is 9 bp. The SAGE program then builds a project database of all possible tags (4^9 = 262,144). The concatemer sequences can then be added. Finally a SAGE report can be generated: this lists the SAGE tags present with their absolute and percentage abundance, accompanied, where requested, by a list of matching genes from the SAGE GenBank Database(s). The SAGE report is saved in Microsoft Access database format for further numerical and statistical analysis with the SAGE program itself, Access and Excel.

It would be an understatement to say that the first SAGE program was not user-friendly and frequently produced error messages, but much improved versions became available during the course of the project. The format of the sequence files proved to be important. The output files from the ABI sequencer included annotation which interfered with the SAGE program. Spaces, for example after every tenth nucleotide, and numbers, for example after every 50th nucleotide, were present in the sequence files. These were not recognised and removed by the SAGE program but ended up within the tags, causing them either to be excluded from analysis or to be assigned an incorrect tag number. Since, at the time, it was not possible to obtain the .seq output files from the ABI sequencer directly in so-called plain format (in which annotation is absent), Dr Keith Vass kindly wrote a short program which operated on the local UNIX computer server within the GCG program package and which could be used to convert multiple sequence files simultaneously to plain format as required by the SAGE program. Nevertheless, this still required multiple file transfers from the PC to UNIX servers and back again.

The GenBank databases are large and unwieldy. The files which I used were the GenBank primate databases, containing only known and characterised genes, not ESTs. These are updated every two months, and took up only two files when I started, but now occupy nine. The files were initially obtained directly from NCBI in hard-copy CD-ROM format but soon thereafter became available for direct on-line file transfer by FTP (file transfer protocol). Unfortunately, the ever-increasing number of GenBank sequences means that three separate SAGE gene databases are now required, rather than the single gene database which was used initially, and each of these three needs to be linked separately to the SAGE files for analysis of each project.

Initially, I verified all putative tag-to-gene matches produced by the SAGE program manually by checking the reference mRNA/cDNA sequence for each gene. This was usually undertaken via the UniGene database, which was designed to bin all transcript sequences from a gene into a single cluster, facilitating transcript profiling. Figure 3.8 shows the entry for the UniGene cluster for gastrin. The last (3'-most) CATG tetranucleotide was identified and the immediately adjacent, down-stream, SAGE tag was confirmed. Obviously, ensuring that the CATG recognition site of *Nla*III is truly the most 3' requires that the mRNA/cDNA sequence contains a polyadenylation site (AAU/TAAA) down-stream. No incorrect matches generated by the SAGE program were ever identified. The UniGene database entries also provide a great deal of additional useful information on each gene, in terms of mRNA expression, chromosomal location and so on, as shown in Figure 3.8.

The SAGE program linked to the down-loaded and personally created GenBank databases was the only mechanism available at the start of this project for matching tags to genes. The advent of NCBI's on-line SAGEmap database has helped greatly, because it provides access not only to curated mRNA/gene sequences, but also to the EST databases, which are far too large to down-load and use locally. SAGEmap can thus yield new, additional tag-to-gene matches, but unfortunately these may not always be correct. For example, some putative matching tags have turned out in the curated mRNA sequences to be located more 5' than expected or in the reverse orientation. Matching genes resulting only from SAGEmap but not from the local SAGE databases therefore always require vigorous checking, although obviously some simply originate from uncharacterised ESTs. As an example, Figure 3.9 shows the SAGEmap entry for the tag matching to gastrin.

Using the SAGE program, analysis of individual SAGE libraries through generating a report was thus eventually a fairly simple task. However, subsequent comparison of SAGE libraries proved to be rather complex. The initial comparisons in the SAGE program were straightforward but these did not produce normalised data, which are required for assessment of fold-differences in tag numbers. Normalisation had to be performed later on the resulting database file using the Excel program. Statistical analysis of the comparative SAGE database files was also simple, but integration of this data with the absolute and normalised tag numbers for each library and with the gene matches required use of, and developing familiarity with, the Microsoft Access program.

3.1.4 Length of SAGE tags: 9 or 10 bp?

The only other problem related to the size of the tags. During SAGE, the tag length is determined at Step 2.2.3.1.6, when the cDNA tags are created and released from the magnetic beads by digestion with the tagging enzyme *BsmF*I. When the project started, the then-current SAGE protocol (version b) stipulated an incubation temperature of 37°C, which yielded tags predominantly of 9 bp length. Subsequent Johns Hopkins protocols (version c onwards) also reported this and recommended incubation instead at the higher temperature of 65°C in order to produce longer tags of 10 bp length or more. This temperature is now also recommended, but was not previously, on the product datasheet from New England Biolabs, as shown in Figure 3.10.

This is because *BsmF*I cleaves slightly differently under different conditions. At 65°C, the enzyme cuts at 10/14 bp from its recognition site GGGAC, to yield a sticky-end (Figure 3.10), whereas the figure is 9/13 at 37°C. During SAGE, the digestion with *BsmF*I is followed by blunt-ending with a DNA polymerase. Even at 37°C, this should yield a 13 bp fragment of which 10 bp should be the SAGE tag: the three nucleotides immediately beside the GGGAC of the *BsmF*I recognition site are actually ATG and form three-quarters of the recognition site of the *Nla*III anchoring enzyme. However, it may be that some of the terminal nucleotides undergo hydrolysis during these manipulations, resulting in short tags.

The SAGE library which was created first, from Tumour 1, thus contained tags mainly of 9 bp. Although the later libraries were created using the higher incubation temperature and therefore contained longer tags, all three libraries were thus analysed in terms of 9 bp tags. Comparisons between my own gastric libraries were straightforward but all other libraries down-loaded for comparison had to be re-analysed in terms of 9 bp tags rather than the usual 10 bp. Further investigation using the on-line SAGEmap tag finder was also affected because the default is 10 bp tags, but these can be achieved simply by adding an extra A, C, G or T onto the 3' end of the 9 bp tag.

3.2 Discussion

Serial analysis of gene expression (SAGE^{TM}) is a patented large-scale mRNA profiling technology which produces comprehensive, quantitative and reproducible gene expression profiles but which is challenging to perform and interpret.

Like most other researchers, I found that SAGE was difficult and time-consuming to set up. SAGE is undoubtedly a valuable technique. However, although the individual stages are simple and involve well-established and relatively straightforward molecular biological techniques, their combination makes the method complex, and it proved to be tricky to set up. As has been clearly demonstrated by the large body of literature subsequently published on technical aspects of SAGE (individually referenced below and discussed in (Yamamoto *et al.* 2001)), the problem was not limited to this laboratory, where I was the first person to establish the method. Further evidence for the difficulty of SAGE is the long time-lag in publications using the technique. Very few papers emerged, except from the originating Johns Hopkins laboratories, for the first few years after SAGE was described and it has only really been from 2000 onwards that primary research papers have emerged, in contrast with reviews, of which there have been many discussing SAGE ever since its first description. Indeed, since the initial report in 1995 by Velculescu *et al*, a number of conferences and workshops focussed entirely on SAGE and its methodology have been held due to popular demand, of which I attended the first, in Amsterdam in 1999.

Part of the difficulty in performing SAGE was the requirement for large amounts of high quality starting RNA. The original Johns Hopkins protocol requires a large quantity of input material: ideally, at least 2.5 μg mRNA, broadly equivalent to 250 μg total RNA, 250 mg tissue or 2.5 x 10^7 cultured cells (technical notes, web-sites of Ambion and Qiagen). The protocol therefore cannot be used to generate expression profiles where RNA is limited, for example, from small tissue biopsies. The RNA must also be of high quality, but unfortunately RNA is labile and readily degraded enzymatically by ribonucleases (Simpson 1987) as well as by adverse physical and chemical conditions.

Extraction of RNA may thus be tricky even from cultured cells, the environment of which can be controlled by the researcher, but locally this procedure worked well. However, RNA purification from clinical material (that is, from patients), was much more difficult, as I found to my cost, in time and effort. Because RNA preparation from cells was successful, the problem with the clinical samples is unlikely to be due to contamination during extraction with exogenous ribonucleases.

Instead, the problem is highly likely to be due to endogenous ribonucleases. This may relate partly to the specific tissue of origin. It is well-recognised that the pancreas and spleen are rich in ribonucleases (technical literature on the web-site of Ambion, a company which specialises in working with RNA). Although published literature on the levels of ribonucleases in gastric tissue is scanty, their presence in abundance might be predicted from the normal digestive functions of the stomach (see Section 1.2.1), which has previously been described as an "intractable tissue source rich in degradative enzymes" (Simpson 1987). Scientists with Ambion also report problems in extracting RNA from stomach. They do not know whether this relates to a high nuclease content or to poor samples (personal communication, Nicola Parsons, Ambion) but this is at least consistent with my findings. In addition, ribonucleases are also activated during and after surgery through tissue damage and death. The solution to obtaining good quality RNA for SAGE (or for other methods of studying gene expression) from clinical material, beyond the routine precautions taken to avoid ribonuclease contamination, is therefore probably to retrieve, dissect and sample tissue from the surgical specimen as swiftly as possible. Nevertheless, as I also found, an analytical agarose gel of the total RNA sample remains mandatory to verify RNA quality and quantity prior to further experiments. In addition, commercial products have since become available which are intended to prevent or minimise RNA degradation in clinical samples during transport and initial storage, such as RNAlater from Ambion. I have no personal experience of using these solutions, but they may well be helpful.

The original requirement for large amounts of high quality RNA has stimulated the development of new protocols using much less starting material. Various technical modifications now enable SAGE to be applied to much smaller quantities of RNA: at least 100-fold, and possibly up to 5000-fold, less may be needed (Velculescu *et al.* 2000). Since my SAGE libraries had already been generated by the time these adaptations were described, I did not use them, but they probably now represent the methods of choice. SADE (a SAGE Adaptation for Downsized Extracts) uses Dynal's oligo(dT)-coated magnetic beads to capture polyA+ mRNA directly from the total RNA or cell lysate (Virlon *et al.* 1999). This procedure substitutes for mRNA purification then cDNA synthesis with biotinylated oligo(dT) followed by capture onto streptavidin-coated Dynabeads. All of the steps from mRNA isolation through to tag release are thus performed directly on the beads, which significantly reduces sample loss. Before managing to get SAGE proper working, I had independently thought of and tried this modification, which is not surprising since the beads were already commercially available from and advertised by Dynal. Unfortunately, however, no ditags resulted after PCR and the attempt was abandoned. Oligo(dT), in the form of a coating inside microcentrifuge tubes (Roche's Streptavidin-Coated Tubes), is used to similar effect in the other adaptations microSAGE (Datson *et al.* 1999) and miniSAGE (Ye *et al.* 2000).

Further modifications include additional PCR steps. In SADE and microSAGE, the ditags generated by the first round of large-scale PCR amplification are re-amplified using extra PCR cycles (Datson *et al.* 1999; Virlon *et al.* 1999). In contrast, SAGE-Lite (Peters *et al.* 1999) and PCR-SAGE (Neilson *et al.* 2000) have adapted Clontech's SMART™ system to generate PCR-amplified cDNA, to increase the amount of input material before proceeding to SAGE proper.

Other technical modifications have improved the efficiency of intermediary SAGE reactions. Most of these adaptations were tested, and in some cases incorporated, in this project. They have been described and, where appropriate, discussed in detail, in the preceding Sections 2.2.3.1 and 3.1.2. The adaptations include the use of biotinylated primers for PCR of the 102 bp ditags to obtain purer 22-26 bp ditag preparations (Powell 1998) and heating the 22-26 bp ditag ligation reaction prior to gel separation of the concatemers (Angelastro *et al.* 2000), both with the aim of obtaining longer cloned concatemer sequences.

Qiagen spin columns have been used by myself and other researchers in preference to phenol/chloroform extraction and ethanol precipitation in various SAGE steps. Angelastro *et al* found that such columns not only are faster and simpler but also remove impurities and improve the efficiency of subsequent enzymatic reactions (Angelastro *et al.* 2000).

More recently, it has been shown that the 22-26 bp ditags are vulnerable to denaturation under low-salt conditions or with slight increases in temperature (Margulies *et al.* 2001). AT-rich ditags have weaker intermolecular bonds and thus lower melting temperatures, so are more likely to denature, which may result in a GC-content bias. It has therefore been recommended that the the 22-26 bp ditags are kept not at room temperature but are instead kept on ice and centrifuged at 4°C (Margulies *et al.* 2001).

The technical difficulty involved in setting up SAGE is also reflected in the recent release of a full commercial SAGE kit and in the increasing detail in and size of the SAGE protocols. Invitrogen now sell a kit called I-SAGE™ which provides all of the numerous reagents required in a high quality form and uses Virlon's adaptation of Dynal's oligo dT-coated magnetic beads (Virlon *et al.* 1999). Although the RNA must still be of high quality, the amount required is around 10 to 100-fold less than for original SAGE protocol. SAGE's complexity is also shown by the increasing length of the various versions of the Johns Hopkins protocol, especially its trouble-shooting section: the protocol is now a total of 27 pages long. Invitrogen's I-SAGE^{TM} manual contains 73 pages, but includes many novel and extremely useful verification steps to check the success of each stage of SAGE.

The software available for analysing SAGE data continues to improve. The SAGE software from the Johns Hopkins laboratory has been regularly updated. Alternative programs for the analysis of tags have been developed, including eSAGE (Margulies *et al.* 2000) and USAGE (van Kampen *et al.* 2000), plus ExProView (Larsson *et al.* 2000) which provides direct visualisation of results. Of these, I have tried only eSAGE but since its statistical analysis produced some results which were clearly incorrect, with p-values over 1, for example, I did not persist with it. The statistical basis for designing and analysing SAGE experiments has also been investigated and discussed in detail (Audic *et al.* 1997; Kal *et al.* 1999; Man *et al.* 2000; Stollberg *et al.* 2000).

Once the SAGE libraries have been produced and analysed, individual SAGE tags may be selected for further study. The National Center for Biotechnology Information (NCBI)'s web-based bioinformatics facilities continue to be extremely useful (Wheeler *et al.* 2001). In particular, the relatively recent SAGEmap database both enables tag-to-gene matching across curated sequences and EST databases and permits the investigation of the expression of individual tags across numerous publicly available SAGE libraries (Lal *et al.* 1999). Unigene remains a most valuable first port-of-call for information about individual transcripts (Wheeler *et al.* 2001). The developing Gene Ontology (the word is derived from Greek and literally means study of being) databases provide a further resource (Ashburner *et al.* 2000). Most genes specifying core biological functions are shared by all eukaryotes and knowledge of these proteins can often be transferred between organisms. Three ontology datasets are being constucted and made available on the World Wide Web: biological process, molecular function and cellular component, which should ease the study of, in particular, less well-characterised genes (Ashburner *et al.* 2000).

Tags either lacking a matching gene or with an incorrect match may be generated in error by SAGE. Occasional sequencing errors are inevitable during SAGE (Stollberg *et al.* 2000). SAGE results usually include a table which lists, amongst other library creation statistics, the absolute number of unique tags obtained (see next chapter, Table 4.1). This is usually regarded as a slight over-estimate of the true number of different genes expressed and a correction which removes around 7% of tags is sometimes applied (Velculescu *et al.* 1995; Zhang, L *et al.* 1997; Velculescu *et al.* 2000), although obviously the figure varies according to the fidelity of the sequencing service used. Since such errors tend to be random, the tags expressed at high numbers are unlikely to be significantly affected; tags present only once are more suspect.

A further source of potential problems arises when the ditags are split by the SAGE program into two tags. The assumption is that the tags are of equal length. However, if, say, one tag had been 10 bp long and the other 8 bp, then splitting an 18 bp (22 bp effectively, taking two nucleotides from the CATG sites on either side) ditag would lead to one correct tag and one incorrect one, with the last nucleotide being erroneous. This may explain some of the findings in the next chapter, where there are multiple tags, differing only in the last nucleotide, for some of the high abundance genes (see next chapter, Table 4.2, e.g. for gastrin). I have not seen this problem identified or discussed elsewhere.

Lastly, the use *in vivo* of alternative polyadenylation cleavage sites downstream of a single polyadenylation signal may lead to more than one SAGE tag from what is in essence a single mRNA species. A recent paper calculated that 2.8% of human transcripts show two or more different SAGE tags corresponding to a single gene because of alternative cleavage site selection alone. Other forms of variant processing such as alternative exon splicing creates similar difficulties.

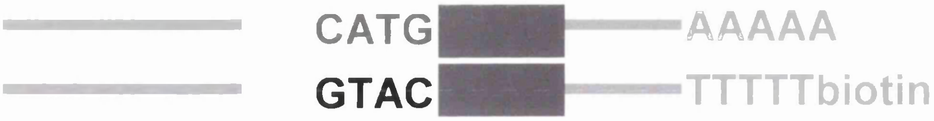
Further information on tags lacking a matching gene is provided by methods based on RT-PCR. Investigation of individual tags is relatively straightforward where the tag clearly corresponds to one gene but may be more difficult where either no matching gene or multiple matches exist. This problem can be addressed by RT-PCR using the short SAGE tag as a primer (Matsumura *et al.* 1999; van den Berg *et al.* 1999; Chen *et al.* 2000). This generates longer, more specific, 3' cDNA fragments which facilitate investigation of the gene, initially through sequencing and further on-line analysis, and which can also be used to check whether the tag is truly differentially expressed between samples of interest (van den Berg *et al.* 1999).

Conclusion. SAGE is an excellent method of large-scale mRNA expression profiling. Although the initial effort involved in setting up SAGE was considerable, requiring time, technical expertise and large amounts of high quality input RNA, the result has been a robust and reproducible technology, and the experience of establishing SAGE has given me an in-depth training in molecular biology and bioinformatics. SAGE produces libraries which are extremely valuable, providing data which are truly comprehensive and quantitative, and which enable the identification of novel genes, as I hope the next chapters will demonstrate: Chapter 4 describes the gene expression profiles of gastric carcinoma and normal stomach created by SAGE; and in Chapter 5 a novel gene which is highly expressed in the stomach is characterised in detail.

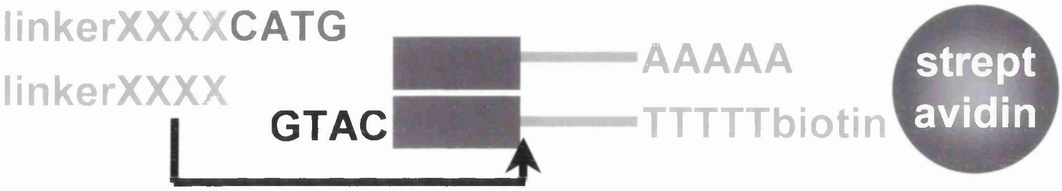
Figure 3.1: Schematic diagram of serial analysis of gene expression (SAGE)

This diagram provides a simplified explanation of the SAGE method (Section 2.2.3).

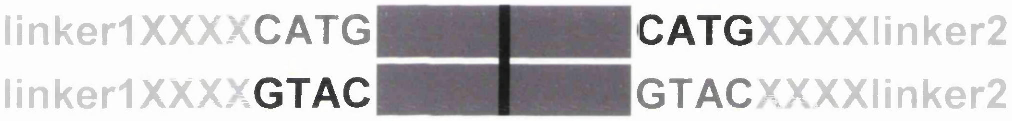
- mRNA → biotinylated cDNA



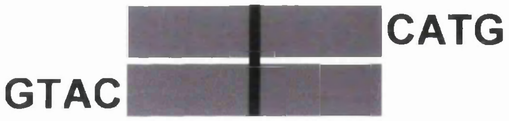
- NlaIII digest (CATG), bind to beads, ligate to linker then BsmFI digest (XXXX)



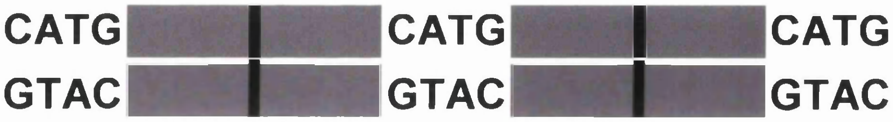
- Ligate linker pairs then PCR



- NlaIII digest to give ditags and purify



- Ligate ditags, purify concatemers then clone



- Sequence clones then use computer software and Genbank databases to analyse tags

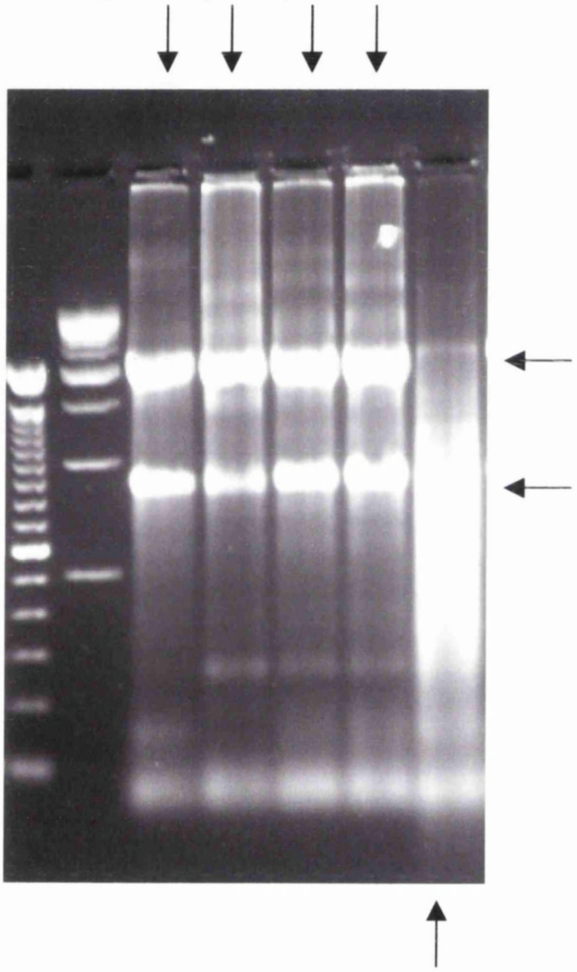
Figure 3.2: Gel electrophoresis of total RNA

The quality (and, to a lesser extent, yield) of total RNA was assessed by electrophoresis on a 1% agarose gel (Section 2.2.2.6), of which an example is shown here. The two lanes on the left hand side contain DNA molecular weight markers (100 bp ladder and λHindIII).

The total RNA run in the next four lanes contains two strong, distinct bands of ribosomal RNA visible at around 4718 and 1847 nucleotides, which represent the 28S and 18S subunits. The light background smearing includes mRNA, which constitutes only 1-5% of total RNA. The low molecular weight bands comprise transfer RNA.

The total RNA run in the lane on the right hand side lacks the two distinct ribosomal bands but instead shows significant heavy smearing which indicates RNA degradation: this sample should therefore not be used for further analysis.

good quality total RNA



ribosomal RNA band at **4.7 kb**

ribosomal RNA band at **1.8 kb**

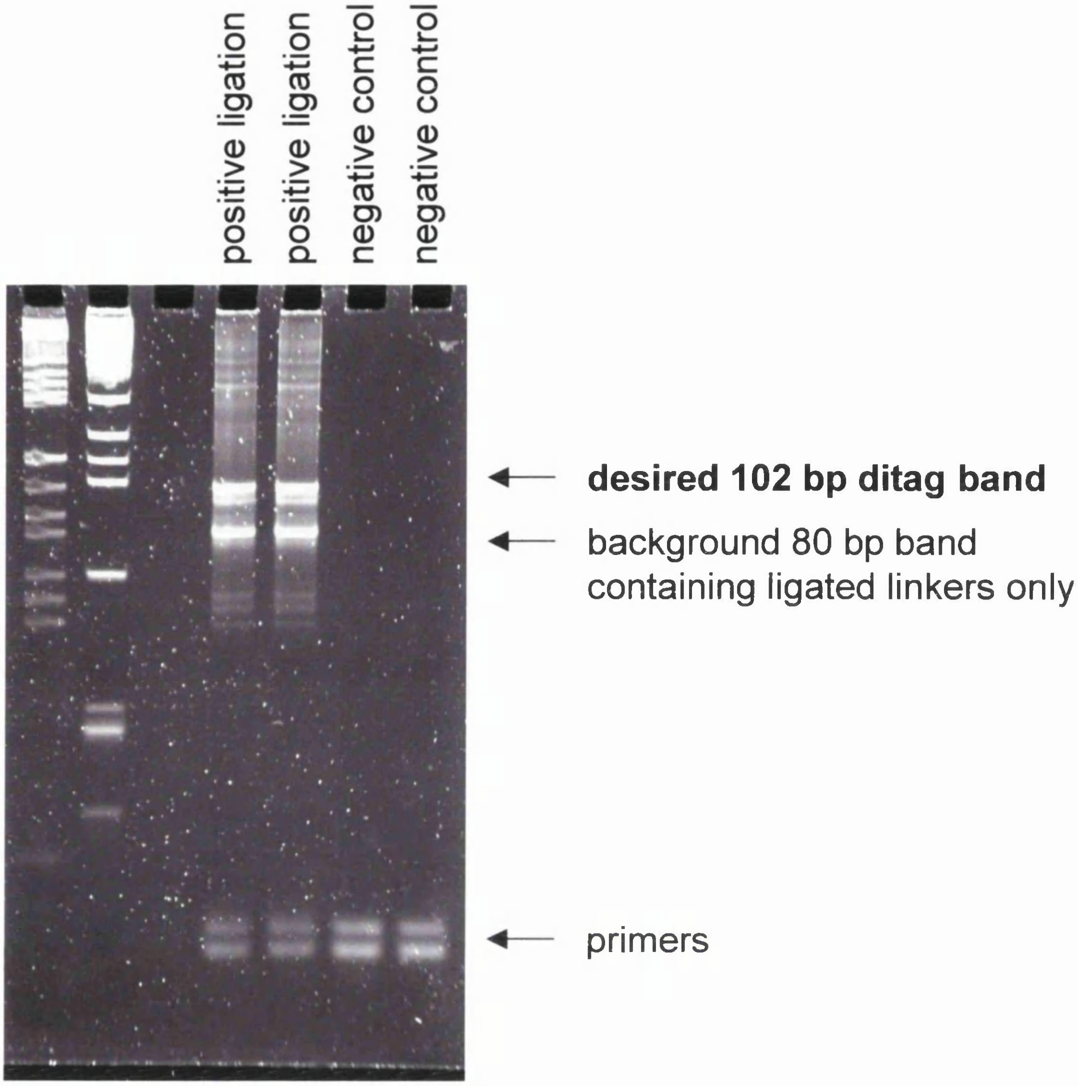
degraded total RNA

Figure 3.3: Gel electrophoresis of small- and large-scale PCRs of 102 bp ditags

Mid-way through the SAGE procedure there is a PCR step which should result in 102 bp ditag PCR products (section 2.2.3.1.9). These contain a central SAGE ditag of 22-26 bp flanked on both sides by linker oligonucleotides of 40 bp. The PCR usually also produces a background band which runs at 80 bp and comprises ligated linker pairs only. The PCR products are analysed and isolated by electrophoresis on 12% polyacrylamide gels, of which examples are shown here. The two lanes on the left hand side contain DNA molecular weight markers (Markers V and VIII).

The upper picture shows the initial small-scale, analytical, gel. This is used to check whether the 102 bp ditag is present at all in the positive ligations (current test SAGE reaction and previous positive SAGE reaction). The two negative control lanes (ligation reaction without ligase enzyme, and no DNA) are used to exclude prior PCR cross-contamination and must lack a PCR product.

The lower picture shows the gel of a large-scale PCR preparation (approximately 100 PCR reactions loaded per gel). The 102 bp band is excised and fragmented then its ditag DNA is purified for further downstream manipulation. The PCR reactions had been purified using Qiagen columns before loading on the gel, hence the absence of lower molecular weight species such as primers.


positive ligation
positive ligation
negative control
negative control
desired 102 bp ditag band
background 80 bp band
containing ligated linkers only
primers

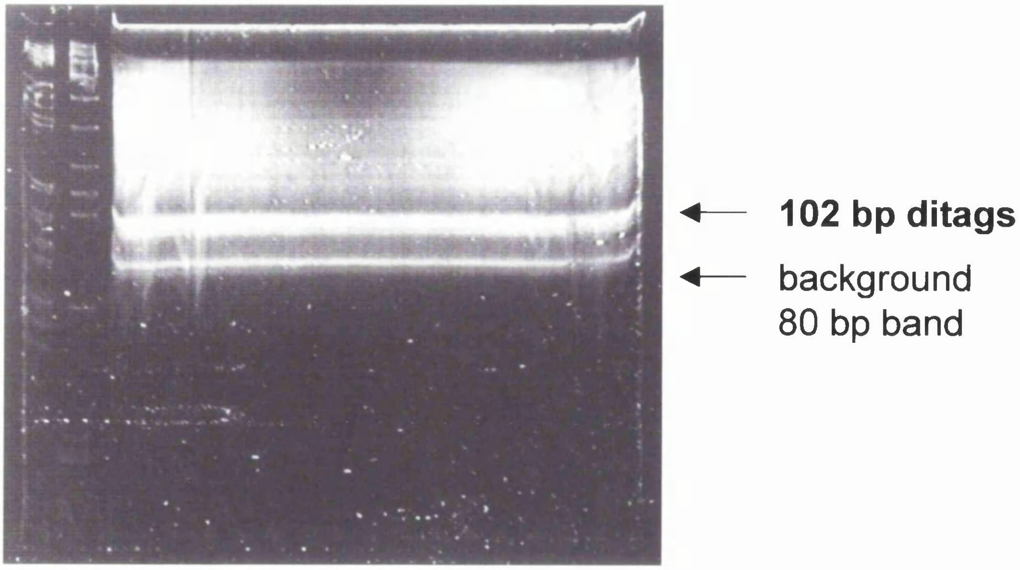
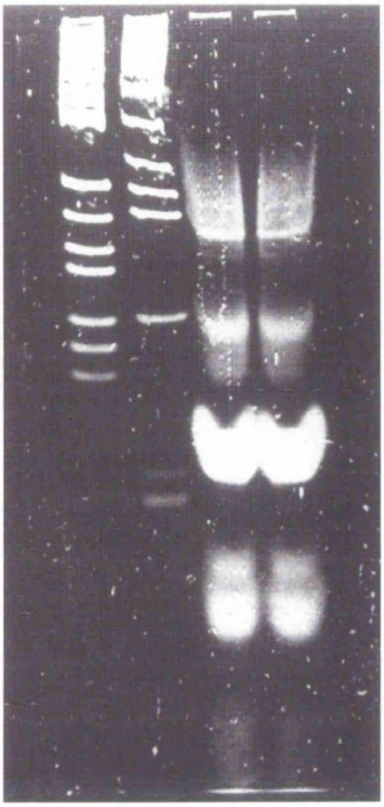

102 bp ditags
background
80 bp band

Figure 3.4: Gel electrophoresis of 26 bp ditags and of concatemer ligation

The large-scale PCR preparation of the 102 bp SAGE ditag is digested with *Nla*III. This releases the 22-26 bp SAGE ditag, leaving the residual 40 bp oligonucleotide linkers, plus partially and completely undigested larger species. These products were separated by 12% PAGE, of which an example is shown here, above. The two lanes on the left hand side contain DNA molecular weight markers (Markers V and VIII).

The band of 22-26 bp SAGE ditags is then excised and purified. The short ditags are then ligated, with the aim of producing long concatemers. These are separated by 8% PAGE, of which an example is shown here, below. The four lanes on the left contain DNA molecular weight markers (Markers V and VIII plus 1 kb and 100 bp ladders). The smeared concatemer DNA which is over 500 bp in size is excised and purified for subsequent cloning.

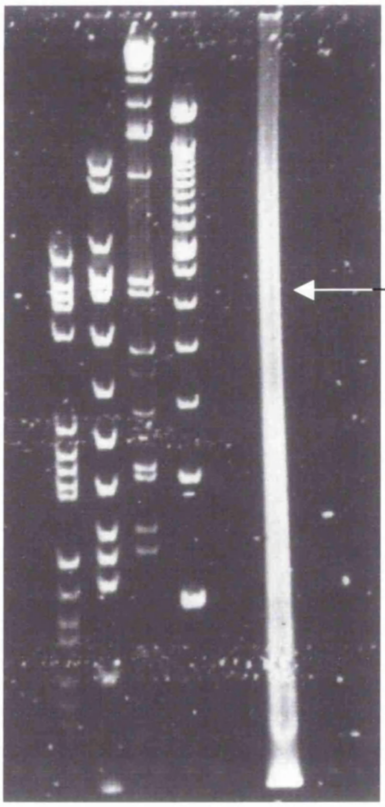


undigested 102 bp ditags

partially digested single linker plus 26 bp ditag

released linkers

26 bp ditags: for excision from gel & purification

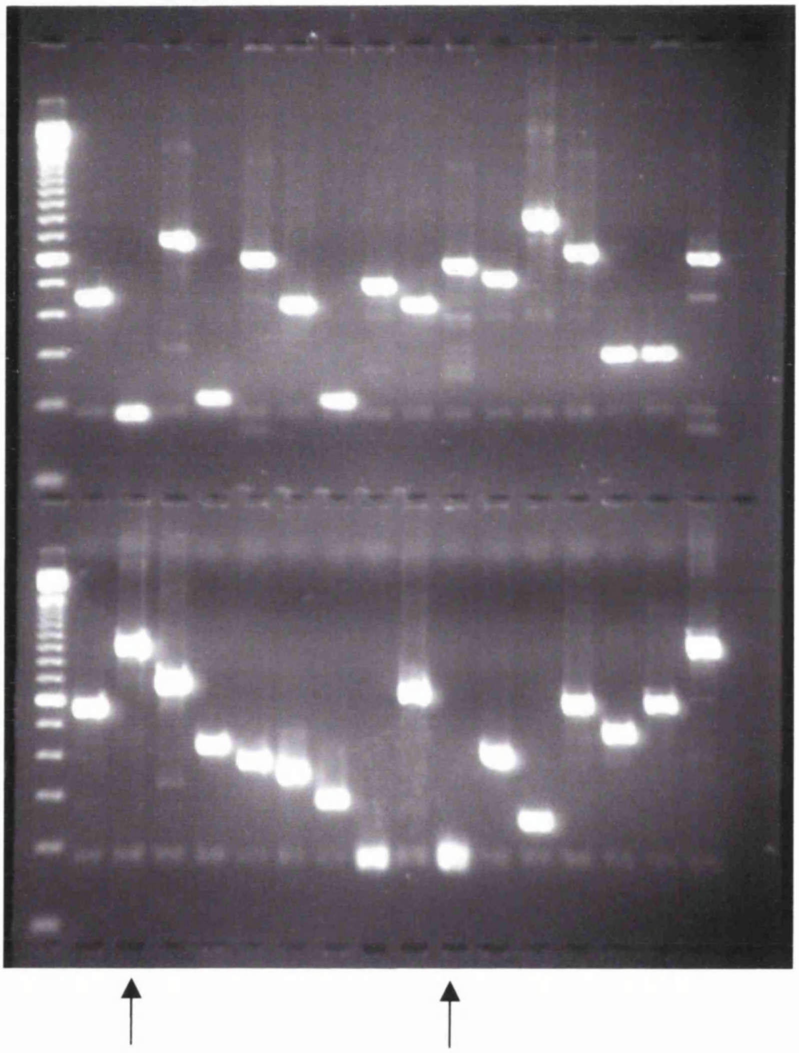


long smear of concatemers (ligated 26 bp ditags): portion > 500 bp for excision from gel & purification

Figure 3.5: Gel electrophoresis of PCR products of cloned concatemers

Single bacterial colonies containing the cloned concatemers are picked and subjected to PCR amplification with primers specific to the vector on either side of the polylinker. The products are separated by electrophoresis on a 2% agarose gel, of which an example is shown here. The lane on the left hand side contains a DNA molecular weight marker (100 bp ladder).

The aim is to achieve PCR products in most lanes of variable size but of at least 500 bp, since these should contain at least 15 tags (around 200 bp of flanking vector plus 12-13 bp per tag). Some colonies inevitably lack inserts, resulting in a 200 bp PCR product. Others often contain only short concatemer inserts. The longer PCR products are then purified for sequencing.



desired PCR product of **vector with long cloned concatemer insert:** here, 900 bp

non-desired PCR product of **background vector without insert:** around 200 bp

Figure 3.6: Sequence of cloned concatemer containing ditags separated by CATGs

This is an example of sequencing of the insert sequence. The insert is delineated by the first and last CATG sites. At the start and the end, there is vector sequence; and the last segment, which contains mostly Ns, represents the end of the sequence. The insert itself is composed of ditags joined serially and separated by the CATG tetranucleotide. At position 481, the expected CATG is lacking but its expected place contains CTTG which presumably represents sequencing error.

When this sequence is put into the SAGE program, the ditags are identified, extracted and split into tags (simply by dividing the ditags down the middle). A catalogue of all tags present is then assembled.

```
ACACTATAGAATACTCAAGCTTGCATGACCTTGGATTTTTTACATCATGC 50
AGGCTGGAGGGAAAGGGGCATGGCATCAAAAAGATTATTCCATGCCATTG 100
TACTCGTGTCGTCCATGCTGGCCCTCGGGGTTTCACCATGAGGTCAGGAG 150
ATTACAGGCATGCCACTGCACTCCCAGGTTCATGGTGGTGCACACTCTGT 200
CGCCCATGACGCAGGGAGAATACGTACCATGGTGAAACCCTGTTTGTTCC 250
CATGTCCCCGTACAGAAGAGCACTCATGGCGGAGGTGGAGTTTCGCCATG 300
ACGCAGGGAGTGCAATGGCATGATGTAAAAAATTAGCCAGGCATGGCAAA 350
ACCCCGCACGTGTACATGGTTGTGGTTAATTTTACATCATGTGAGATCCA 400
GAATACGTACCATGTCCATTAAGCCTAGATGTGCATGGTACCGTATTCTC 450
CCCTGGCATGATCGGGCCCGGAGTTTTACCCTTGAATTAAAATTAAGGCT 500
GTTCATGCCCATCGTCCTCAGCCTTCCATGCCCCTTGCAGGTCCAGCACC 550
ATGCTTCCCCAGCACCGGTTGGCATGCCACCGACTTTTCGATTCATGCCT 600
GCAGGTCGACTCTAAGGATCCCCGGTACCGAGCTCGAATTCCCCTATAGN 650
GAGTCGGNTTACAATCNCTGGCCGGGTTTTTNAAANTNNNNNTNNNNNCN 700
NNGGTANNTNTTTTCNCTNNNNNTTNNTTNNNTCNTNNNNTNNTNTNNNT 750
NTNNNNNNTTNTCNNTNTNCCTTTCNNNNNCNNNTNTNNNNNNNNNNNNT 800
CNTNCNNNCTNNNTNNNTNTNNNTNNNNNTTTNNNNCNNTNCTNNT 846
```

Figure 3.7: SAGE project report linked to genetic database

This is an example of a SAGE tag report including gene matches. The `BOLD HEADINGS` have been added to ease interpretation. For SAGE project files, first the parameters are specified: the anchoring enzyme is *Nla*III, with its CATG recognition site; the tag length is 9 bp; and the maximum ditag length, over which size any ditags are excluded from analysis, is 28 bp. The SAGE program then builds a project database of all possible tags (4^9 = 262,144). Thereafter, the concatemer sequences can be added. Thus the SAGE project has been created and used to form a SAGE report: in this case, the project contains 987 sequences (files). 1806 duplicate ditags have been identified and excluded from further analysis. Other "excluded tags" comprise mainly sequences originating from the linker oligonucleotides. 8433 tags remain for full analysis in this report, which is saved not only as a text file but also as in Microsoft Access database format, which can then be used for further analyses such as comparisons with other SAGE libraries or calculations of statistical significance.

The tags are then listed, ranked by absolute tag count. For this example report, only tags present at an absolute abundance of 30 or more are included. Ordinarily, more tags would be reported, eventually down to one tag only. The overall % abundance is also listed. The "Tag BaseFour Number" is simply the number in the underlying tag database, where AAAAAAAAA would be 1 and TTTTTTTTT would be 4^9.

The report has been linked to a gene database. The gene databases are created separately from down-loaded Genbank files. When this report was created, the Genbank files used were the nine primate gene databases; though these should contain only characterised sequences not ESTs, they fill three separate Access databases, each of which needs to be linked separately to the SAGE files for analysis of each project. Thus the linked genes listed here represent only around one-third of all possibilities.

The number of matches to this tag in the entire database is indicated as "Tags Noted". The number of tags identified at the 3' end of the sequence entry is indicated as "Tags Collected," since in cDNA mode, as used here, only these can be SAGE tags. "Trailer" means the six adjacent nucleotides immediately 3' to the tag, which, if the tag is longer than 9 bp, can be used better to identify the gene where there are multiple matches. The tag entry for gastrin, which is used as an example throughout this thesis, is highlighted. Class A tags were annotated in GenBank as mRNAs, Class B tags as ESTs, and Class C tags were not listed as either A or B.

DETAILS OF SAGE REPORT

```
Date: 06-28-2001
Report File Name = N:\sage new\sage second projects\sage normal try2\rep-13.rpt
```

DETAILS OF SAGE PROJECT BEING REPORTED

```
Project # 1
Project File = N:\sage new\sage second projects\sage normal try2\try2.sum
SearchName = try2
Anchoring Enzyme = NlaIII - CATG
Tag Length =  9
DiTag Length =  28
Total Files =  987
Total Tags =  8433
Total Duplicate Dimers =  1806

Total tags in selected Projects =  8433
```

MICROSOFT ACCESS FILE TO WHICH SAGE REPORT RESULTS BEING SAVED

```
Tag Abundance Report
MS Access File Name = C:\PROGRAM FILES\SAGE300\SAGE.MDB
```

LIST OF TAGS WHICH IF PRESENT WOULD BE EXCLUDED FROM FINAL ANALYSIS

```
Excluded Tags
Tag Sequence  Tag Base Four Number
CCTATTAA       875761
CCCGTACA       874181
CCCTATTAA      89329
CCCGTACA       612037
CCTATTAA       351473
CCCGTACA       349893
CCTATTAA       613617
CCCCGTACA      87749
CTATTAAG       357315
CCGTACAT       350996
CTATTAAG       816067
CCCGTACA       1005253
CCTATTAA       941297
CCCGTACA       939717
CCTATTAA       1006833
CCCGTACA       808645
CTATTAAG       881603
CCGTACAT       875284
ACTATTAA       859377
TCCGTACA       906949
GCTATTAA       892145
GCCGTACA       890565
TCTATTAA       908529
ACCGTACA       857797
CATATTAA       871665
CTCGTACA       882373
CGTATTAA       879857
CTTATTAA       883953
CACGTACA       870085
CGCGTACA       878277
CCAATTAA       872689
CCTGTACA       876229
CCCATTAA       873713
CCGGTACA       875205
CCGATTAA       874737
CCAGTACA       873157
CCATTAAG       873411
CCGTACAT       875284

Total tags after excluding tags =  8433
```

NUMERICAL LIST OF TAGS PRESENT (REPORT REQUESTED FOR ONLY THOSE TAGS PRESENT IN PROJECT AT AN ABSOLUTE ABUNDANCE OF 30 OR MORE)

```
Count          Percent        Tag Sequence  Tag BaseFour Number

 659           7.8145         CTCCCCCAA      120145
Genes in Class =  1    *  Tags in Class =  659
Cumulative Gene Count =  1    *  Cumulative Tag Count =  659
```

```
 181            2.1463      AAGGGAGCA      10789
Genes in Class =  1   *  Tags in Class =  181
Cumulative Gene Count =  2    *  Cumulative Tag Count =  840

 143            1.6957      CTGGCCCTC      125278
Genes in Class =  1   *  Tags in Class =  143
Cumulative Gene Count =  3    *  Cumulative Tag Count =  983

 112            1.3281      GCTGGAGGA      162345
Genes in Class =  1   *  Tags in Class =  112
Cumulative Gene Count =  4    *  Cumulative Tag Count =  1095

 101            1.1976      TCCCTATTA      218941
Genes in Class =  1   *  Tags in Class =  101
Cumulative Gene Count =  5    *  Cumulative Tag Count =  1196

 92             1.0909      TCCCCGTAC      218546
Genes in Class =  1   *  Tags in Class =  92
Cumulative Gene Count =  6    *  Cumulative Tag Count =  1288

 85             1.0079      ATGTAAAAA      60417
Genes in Class =  1   *  Tags in Class =  85
Cumulative Gene Count =  7    *  Cumulative Tag Count =  1373

 80             0.9486      GGGCTGGGG      173995
Genes in Class =  1   *  Tags in Class =  80
Cumulative Gene Count =  8    *  Cumulative Tag Count =  1453

 77             0.913       TCATTCTGA      216953
Genes in Class =  1   *  Tags in Class =  77
Cumulative Gene Count =  9    *  Cumulative Tag Count =  1530

 71             0.8419      CCACTGCAC      83858
Genes in Class =  1   *  Tags in Class =  71
Cumulative Gene Count =  10    *  Cumulative Tag Count =  1601

 62             0.7352      CCCATCGTC      86894
Genes in Class =  1   *  Tags in Class =  62
Cumulative Gene Count =  11    *  Cumulative Tag Count =  1663

 54             0.6403      TGTGTTGAG      244707
Genes in Class =  1   *  Tags in Class =  54
Cumulative Gene Count =  12    *  Cumulative Tag Count =  1717

 49             0.581       CGCCGCCGG      104027
Genes in Class =  1   *  Tags in Class =  49
Cumulative Gene Count =  13    *  Cumulative Tag Count =  1766

 47             0.5573      GCCGAGGAA      153761
Genes in Class =  1   *  Tags in Class =  47
Cumulative Gene Count =  14    *  Cumulative Tag Count =  1813

 46             0.5454      CCCGTCCGG      88923
 46             0.5454      GTACGTATT      181968
Genes in Class =  2   *  Tags in Class =  92
Cumulative Gene Count =  16    *  Cumulative Tag Count =  1905

 44             0.5217      GAAATAAAG      131843
Genes in Class =  1   *  Tags in Class =  44
Cumulative Gene Count =  17    *  Cumulative Tag Count =  1949

 42             0.498       AAATCCTGG      3451
 42             0.498       GGCTGGGGG      171691
Genes in Class =  2   *  Tags in Class =  84
Cumulative Gene Count =  19    *  Cumulative Tag Count =  2033

 36             0.4268      GTGAAACCC      188438
 36             0.4268      GTGATCAGC      189258
Genes in Class =  2   *  Tags in Class =  72
Cumulative Gene Count =  21    *  Cumulative Tag Count =  2105

 35             0.415       AAGGTGGAG      11171
Genes in Class =  1   *  Tags in Class =  35
Cumulative Gene Count =  22    *  Cumulative Tag Count =  2140

 34             0.4031      GCGGAGGTG      157871
Genes in Class =  1   *  Tags in Class =  34
Cumulative Gene Count =  23    *  Cumulative Tag Count =  2174
```

```
 32           0.3794       TTCCCCCAA      251217
 32           0.3794       TTGGTCCTC      256862
Genes in Class =  2   *  Tags in Class =  64
Cumulative Gene Count =  25   *  Cumulative Tag Count =  2238

Genes identified in project =  25
```

GENETIC DATABASE TO WHICH THIS SAGE REPORT LINKED

DataBase Link

```
Database = n:\sage new\sage databases\apr01\apr01_9\GENETAG.dum
```

NUMERICAL LIST OF TAGS PRESENT WITH CORRESPONDING DATABASE MATCHES
(REPORT REQUESTED AN ARBITRARY LIMIT OF 10 MATCHING GENE ENTRIES)

Tag Count   Abundance (%) Tag Sequence Tag Base Four Number

```
 659          7.8145       CTCCCCCAA      120145
Noted Tags =  4            Collected Tags =  0

 181          2.1463       AAGGGAGCA      10789
Noted Tags =  45           Collected Tags =  28
L38562, Class A, Trailer CCGTGG, Position  346
      Human clone rev16/20Fab, immunoglobulin lambda light chain mRNA,
X57824, Class A, Trailer CCGTGG, Position  658
      Human rearranged immunoglobulin lambda light chain mRNA.
X57823, Class A, Trailer CCGTGG, Position  658
      Human rearranged immunoglobulin lambda light chain mRNA.
X57822, Class A, Trailer CCGTGG, Position  656
      Human rearranged immunoglobulin lambda light chain mRNA.
X57821, Class A, Trailer CCGTGG, Position  654
      Human rearranged immunoglobulin lambda light chain mRNA.
X57820, Class A, Trailer CCGTGG, Position  687
      Human rearranged immunoglobulin lambda light chain mRNA.
X57818, Class A, Trailer CCGTGG, Position  655
      Human rearranged immunoglobulin lambda light chain mRNA.
X57817, Class A, Trailer CCGTGG, Position  652
      Human rearranged immunoglobulin lambda light chain mRNA.
X57816, Class A, Trailer CCGTGG, Position  668
      Human rearranged immunoglobulin lambda light chain mRNA.
X57815, Class A, Trailer CCGTGG, Position  663
      Human rearranged immunoglobulin lambda light chain mRNA.

 143          1.6957       CTGGCCCTC      125278
Noted Tags =  7            Collected Tags =  2
X52003, Class A, Trailer GGCACC, Position  88
      H.sapiens pS2 protein gene.
X00474, Class A, Trailer GGCACC, Position  88
      Human pS2 mRNA induced by estrogen from human breast cancer cell

 112          1.3281       GCTGGAGGA      162345
Noted Tags =  12           Collected Tags =  2
V00511, Class A, Trailer AGAAGA, Position  241
      Human mRNA encoding pregastrin (a regulatory hormone of gastric
X00183, Class C, Trailer AGAAGA, Position  624
      Human gastrin gene.

 101          1.1976       TCCCTATTA      218941
Noted Tags =  4            Collected Tags =  0

 85           1.0079       ATGTAAAAA      60417
Noted Tags =  7            Collected Tags =  1
AL137528, Class A, Trailer GGAAAA, Position  4064
      Homo sapiens mRNA; cDNA DKFZp434P1818 (from clone DKFZp434P1818);

 80           0.9486       GGGCTGGGG      173995
Noted Tags =  14           Collected Tags =  2
Z49148, Class A, Trailer TCCTCC, Position  579
      H.sapiens mRNA for ribosomal protein L29.
V00488, Class C, Trailer AGGGAG, Position  1020
      Human alpha-globin germ line gene.

 77           0.913        TCATTCTGA      216953
Noted Tags =  4            Collected Tags =  0

 71           0.8419       CCACTGCAC      83858
```

```
Noted Tags =  994           Collected Tags =  28
X83301, Class A, Trailer TCCAGC, Position  1722
       H.sapiens SMA5 mRNA.
X83299, Class A, Trailer TCCAGC, Position  1252
       H.sapiens SMA3 mRNA.
X77626, Class C, Trailer TCCAGC, Position  4
       H.sapiens simple sequence repeat region clone 9q3f10.
Z46632, Class A, Trailer TCCAGC, Position  3422
       H.sapiens HSPDE4C1 gene for 3',5'-cyclic AMP phosphodiesterase.
Z77985, Class C, Trailer TCCAGC, Position  155
       H.sapiens flow-sorted chromosome 6 HindIII fragment, SC6pA6B4.
Z78963, Class C, Trailer TCCAGG, Position  138
       H.sapiens flow-sorted chromosome 6 TaqI fragment, SC6pA1F11.
AL442085, Class A, Trailer TCCACC, Position  2136
       Homo sapiens mRNA; cDNA DKFZp547I084 (from clone DKFZp547I084).
AL390174, Class A, Trailer TCCAGC, Position  3503
       Homo sapiens mRNA; cDNA DKFZp547J184 (from clone DKFZp547J184).
AL353951, Class A, Trailer TCCAGC, Position  2750
       Homo sapiens mRNA; cDNA DKFZp761A0423 (from clone DKFZp761A0423).
AL161994, Class A, Trailer TCCAGC, Position  1405
       Homo sapiens mRNA; cDNA DKFZp564A0772 (from clone DKFZp564A0772);

 62            0.7352       CCCATCGTC      86894
Noted Tags =  6             Collected Tags =  1
X15759, Class A, Trailer CTAGAA, Position  619
       H.sapiens mitochondrial mRNA for cytochrome c oxidase subunit II.

 54            0.6403       TGTGTTGAG      244707
Noted Tags =  8             Collected Tags =  3
X03558, Class A, Trailer AGCTTC, Position  1280
       Human mRNA for elongation factor 1 alpha subunit (EF-1 alpha).
X16872, Class C, Trailer ACCTTC, Position  1347
       Human DNA for elongation factor 1-alpha (clone lambda-9).
X16869, Class A, Trailer AGCTTC, Position  1242
       Human mRNA for elongation factor 1-alpha (clone CEF4).

 47            0.5573       GCCGAGGAA      153761
Noted Tags =  3             Collected Tags =  1
X53505, Class A, Trailer GGCATT, Position  79
       Human mRNA for ribosomal protein S12.

 46            0.5454       CCCGTCCGG      88923
Noted Tags =  1             Collected Tags =  0

 46            0.5454       GTACGTATT      181968
Noted Tags =  1             Collected Tags =  0

 44            0.5217       GAAATAAAG      131843
Noted Tags =  17            Collected Tags =  4
AL137306, Class A, Trailer TCTGTA, Position  1560
       Homo sapiens mRNA; cDNA DKFZp586K1417 (from clone DKFZp586K1417).
X16110 M27, Class C, Trailer CACCCA, Position  2291
       H.sapiens IGHG3 gene for immunoglobulin heavy chain constant region
Y14737, Class A, Trailer CACCCA, Position  1594
       Homo sapiens mRNA for immunoglobulin lambda heavy chain.
Y14735, Class A, Trailer CACCCA, Position  1552
       Homo sapiens mRNA for immunoglobulin kappa heavy chain.

 42            0.498        AAATCCTGG      3451
Noted Tags =  3             Collected Tags =  2
X51698, Class A, Trailer GTTTTC, Position  480
       H.sapiens spasmolytic polypeptide (SP) mRNA.
X97793, Class C, Trailer GTTTTC, Position  181
       H.sapiens SML1 gene exon 4 and flanking regions.

 42            0.498        GGCTGGGGG      171691
Noted Tags =  6             Collected Tags =  0

 36            0.4268       GTGAAACCC      188438
Noted Tags =  4167          Collected Tags =  81
U02068, Class A, Trailer CATCTC, Position  96
       Human clone 9 Alu repeat mRNA sequence.
AJ223075, Class A, Trailer TGTCTC, Position  3470
       Homo sapiens mRNA for TRIP protein.
U03115, Class C, Trailer CATCTC, Position  77601
       Human V beta T-cell receptor (TCRBV) gene locus.
U79652, Class C, Trailer C     , Position  1038
```

```
       Human Treacher Collins syndrome (TCOF1) gene, exons 14, 15 and 16.
Y15724, Class C, Trailer GTCTCT, Position  18345
       Homo sapiens SERCA3 gene, exons 1-7 (and joined CDS).
X65708, Class C, Trailer TGTCTC, Position  6651
       H.sapiens RRM1 gene for ribonucleoside diphosphate reductase M1
Z29096, Class C, Trailer TGTCTC, Position  1931
       H.sapiens (hRPB25) gene for RNA polymerase II subunit exon 1.
AF044967, Class C, Trailer CGTCTC, Position  1209
       Homo sapiens polio virus related protein 2 gene, exons 8 and 9.
AF044962, Class C, Trailer TGTCTC, Position  1255
       Homo sapiens polio virus related protein 2 gene, exon 2.
X71874, Class C, Trailer CGTCTC, Position  13670
       H.sapiens genes for proteasome-like subunit (MECL-1),

 36            0.4268       GTGATCAGC      189258
Noted Tags =  4             Collected Tags =  3
AJ001402, Class A, Trailer TGCCTG, Position  3558
       Homo sapiens mRNA for MUC5AC protein (tracheal), partial.
AJ001403, Class C, Trailer TGCCTG, Position  9085
       Homo sapiens mNRA for MUC5AC protein (placental).
Z48314, Class A, Trailer TGCCTG, Position  3166
       H.sapiens mRNA for apomucin.

 35            0.415        AAGGTGGAG      11171
Noted Tags =  11            Collected Tags =  1
X80822, Class A, Trailer GAGATC, Position  403
       H.sapiens mRNA for ORF.

 34            0.4031       GCGGAGGTG      157871
Noted Tags =  3             Collected Tags =  1
AJ294729, Class A, Trailer GACGGC, Position  1032
       Homo sapiens partial mRNA for immunoglobulin heavy chain constant

 32            0.3794       TTCCCCCAA      251217
Noted Tags =  2             Collected Tags =  0

 32            0.3794       TTGGTCCTC      256862
Noted Tags =  5             Collected Tags =  1
Z12962 S45, Class A, Trailer TGCCCT, Position  348
       H.sapiens mRNA for homologue to yeast ribosomal protein L41.
```

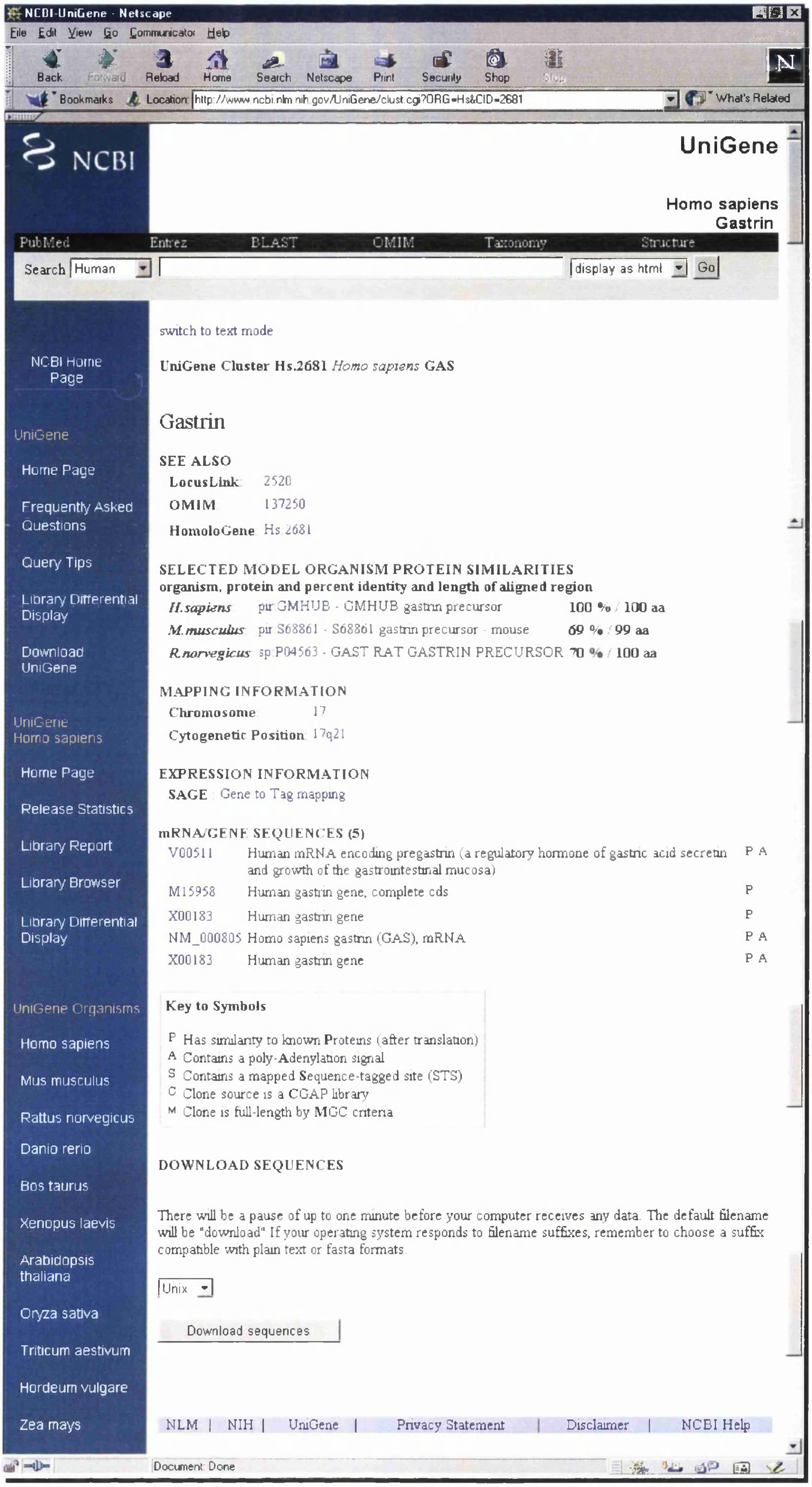
Figure 3.8: UniGene cluster for gastrin

Putative SAGE tag-to-gene matches, and the resulting genes of interest, in this case gastrin, were further investigated through NCBI's UniGene web-site and databases: http://www.ncbi.nlm.nih.gov/UniGene/index.html/.

In this case, the UniGene cluster number is Hs. (Homo sapiens) 2681. This number is arbitrary and occasionally changes, as the clusters are re-assigned, so the number cannot, unlike the gene name itself, be used as a permanent record. The symbol for the gene is GAS. Hyperlinks are provided to other databases: LocusLink; the OMIM (On-Line Mendelian Inheritance in Man) database, which provides further information on numerous individual genes; and HomoloGene, which describes homologous genes in the other model organisms included in the UniGene databases, including mouse, rat and cow amongst others. The latter are then described in more detail in the "Selected Model Organism Protein Similarities."

Information is then provided on the gene's chromosomal location and mRNA expression pattern. The latter is usually generated from the library source(s) of matching ESTs. Curated mRNA/Gene sequences are listed and can be displayed and used to verify a putative SAGE tag-to-gene match. The Genbank Accession Number of the original reference sequence usually starts, as here, with the characters NM_. Within the mRNA/cDNA sequence, the last (3'-most) CATG tetranucleotide is identified and the immediately adjacent, down-stream, SAGE tag confirmed. Obviously, ensuring that the CATG recognition site of *Nla*III is truly the most 3' requires that the mRNA/cDNA sequence contains a polyadenylation site (AAU/TAAA) down-stream, as indicated in this list by the symbol A surrounded by a square.

Normally, these curated sequences would be followed by matching ESTs, but none are listed for gastrin. This is perhaps not surprising since most of the EST libraries are derived from cancers, which would not be expected to contain the highly differentiated antral tissue containing the G-cells from which gastrin is derived (see Section 1.2.2).



Homo sapiens
Gastrin

switch to text mode

UniGene Cluster Hs.2681 *Homo sapiens* **GAS**

Gastrin

SEE ALSO

LocusLink:	2520
OMIM:	137250
HomoloGene:	Hs.2681

SELECTED MODEL ORGANISM PROTEIN SIMILARITIES
organism, protein and percent identity and length of aligned region

H.sapiens:	pir:GMHUB - GMHUB gastrin precursor	100 % / 100 aa
M.musculus:	pir:S68861 - S68861 gastrin precursor - mouse	69 % / 99 aa
R.norvegicus:	sp:P04563 - GAST RAT GASTRIN PRECURSOR	70 % / 100 aa

MAPPING INFORMATION

Chromosome:	17
Cytogenetic Position:	17q21

EXPRESSION INFORMATION
SAGE: Gene to Tag mapping

mRNA/GENE SEQUENCES (5)

V00511	Human mRNA encoding pregastrin (a regulatory hormone of gastric acid secretin and growth of the gastrointestinal mucosa)	P A
M15958	Human gastrin gene, complete cds	P
X00183	Human gastrin gene	P
NM_000805	Homo sapiens gastrin (GAS), mRNA	P A
X00183	Human gastrin gene	P A

Key to Symbols

P Has similarity to known Proteins (after translation)
A Contains a poly-Adenylation signal
S Contains a mapped Sequence-tagged site (STS)
C Clone source is a CGAP library
M Clone is full-length by MGC criteria

DOWNLOAD SEQUENCES

There will be a pause of up to one minute before your computer receives any data. The default filename will be "download". If your operating system responds to filename suffixes, remember to choose a suffix compatible with plain text or fasta formats.

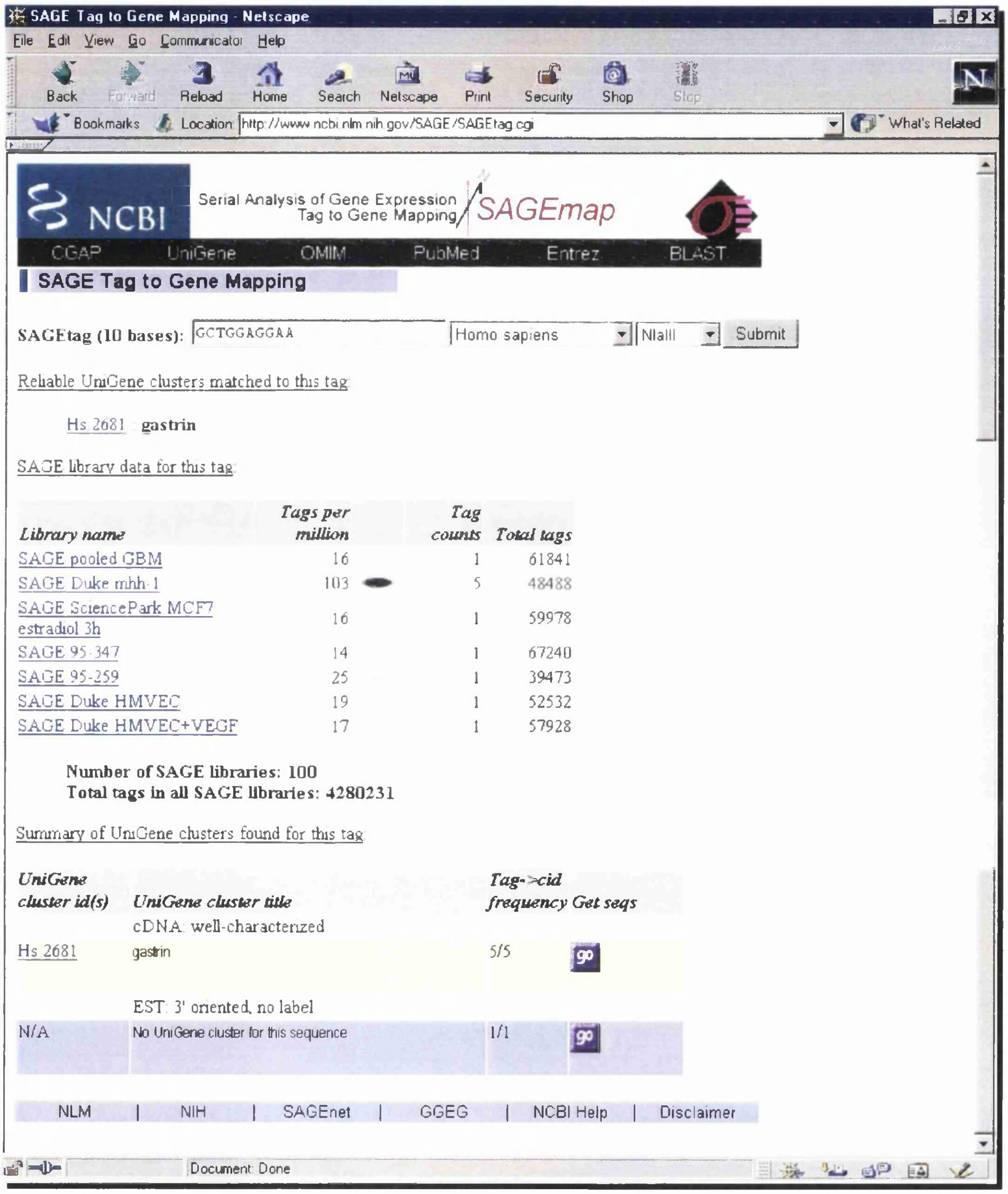
Unix

Download sequences

Figure 3.9: SAGEmap entry for the gastrin tag

Tags of interest resulting from SAGE could also be further investigated through NCBI's SAGEmap web-site: http://www.ncbi.nlm.nih.gov/SAGE/. In this case, the tag matched to gastrin mRNA. The 9 bp tag (GCTGGAGGA), plus a terminal A, C, G or T, is entered into the query box at the top of the page, along with details of the sampled organism and Anchoring Enzyme used. The only reliable UniGene Cluster matching to the tag was for gastrin (see also Figure 3.8). At the bottom of the page is provided a list of all established cDNAs, or ESTs, matching to the tag, plus their orientation (5' or 3'). In the middle of the page, SAGEmap also lists the on-line SAGE libraries in which the tag is present and its absolute and relative abundance therein.

For interest, the single library with a relatively high gastrin tag count (103 per million) is Duke mhh-1, which was derived from a medulloblastoma, that is, a primitive neuro-ectodermal tumour of the cerebellum. This is not surprising: gastrin mRNA has previously been demonstrated in 10 out of 11 medulloblastomas (Schaer *et al.* 1999). These tag levels are nevertheless a great deal lower than those in normal gastric antrum, where the gastrin tag count was 100-fold higher, at around 10,000 per million.


SAGE Tag to Gene Mapping - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.ncbi.nlm.nih.gov/SAGE/SAGEtag.cgi
What's Related
NCBI
Serial Analysis of Gene Expression
Tag to Gene Mapping
SAGEmap
CGAP UniGene OMIM PubMed Entrez BLAST
Document: Done

SAGE Tag to Gene Mapping

SAGEtag (10 bases): GCTGGAGGAA | Homo sapiens | NlaIII | Submit

Reliable UniGene clusters matched to this tag:

Hs.2681 : **gastrin**

SAGE library data for this tag:

Library name	*Tags per million*	*Tag counts*	*Total tags*
SAGE pooled GBM	16	1	61841
SAGE Duke mhh-1	103	5	48488
SAGE SciencePark MCF7 estradiol 3h	16	1	59978
SAGE 95-347	14	1	67240
SAGE 95-259	25	1	39473
SAGE Duke HMVEC	19	1	52532
SAGE Duke HMVEC+VEGF	17	1	57928

Number of SAGE libraries: 100
Total tags in all SAGE libraries: 4280231

Summary of UniGene clusters found for this tag:

UniGene cluster id(s)	*UniGene cluster title*	*Tag->cid frequency*	*Get seqs*
	cDNA: well-characterized		
Hs.2681	gastrin	5/5	go
	EST: 3' oriented, no label		
N/A	No UniGene cluster for this sequence	1/1	go

NLM | NIH | SAGEnet | GGEG | NCBI Help | Disclaimer

Figure 3.10: Datasheet from New England Biolabs on Tagging Enzyme *BsmF*I

During SAGE, the tag length is determined at Step 2.2.3.1.6, when the cDNA tags are created and released from the magnetic beads by digestion with the tagging enzyme *BsmF*I. As shown here, in the product datasheet from New England Biolabs, *BsmF*I cleaves slightly differently under different conditions. At 65°C, the enzyme cuts at 10/14 bp from its recognition site GGGAC, to yield a sticky-end, whereas the figure is 9/13 under other conditions, including the lower incubation temperature of 37°C. The latter was recommended in initial Johns Hopkins' protocols, so SAGE tags generated in earlier libraries, at 37°C, are shorter than those created using the higher temperature of 65°C.

BsmF I

NEB 4 BSA 65°

#R0572S	100 units	**$60 (USA)**
#R0572L	500 units	**$240 (USA)**

5′ ... G G G A C $(N)_{10}$^ ... 3′
3′ ... C C C T G $(N)_{14}$^ ... 5′

Source: *Bacillus stearothermophilus* F (Z. Chen)

Reaction Buffer: (Supplied with enzyme) NEBuffer 4 + BSA
50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate, 1 mM dithiothreitol (pH 7.9 @ 25°C). Supplement with 100 µg/ml BSA Incubate at 65°C.

Ligation and Recutting: After 10-fold overdigestion with BsmF I, > 95% of the DNA fragments can be ligated and recut.

Concentration: 2,000 units/ml. Assayed on pBR322 DNA.

Storage Conditions: 50 mM NaCl, 10 mM Tris-HCl (pH 7.4), 1 mM dithiothreitol, 0.1 mM EDTA, 200 µg/ml BSA, and 50% glycerol. Store at -20°C.

Diluent Compatibility: Diluent A

Heat Inactivation: 80°C for 20 minutes.

Note: Occasionally, BsmF I has been shown to cleave the sequence GGGAC(9/13). The exact frequency of this occurrence has yet to be determined. BsmF I is an isoschizomer of Fin I. Incubation at 37°C results in 50% activity.

4 Large-scale expression profiling of gastric carcinoma and normal stomach by SAGE and its validation

Having described how the method of serial analysis of gene expression was established locally, this chapter now presents the large-scale mRNA profiles of gastric adenocarcinoma and normal gastric antrum obtained using SAGE. The expression patterns of selected candidate genes were validated by Northern blotting and imunohistochemistry.

4.1 Results

4.1.1 Gastric SAGE profiles

4.1.1.1 Generation of gastric SAGE tag libraries

SAGE libraries were created from two gastric adenocarcinomas, of the same subtype (distal and intestinal) but from different patients, and from normal gastric antral mucosa. After excluding inappropriate tags (resulting from duplicate ditags or linker sequences), a total of 29,480 tags, or transcripts, derived from 10866 different genes (unique tags), was obtained. The libraries from Tumour 1, Tumour 2 and normal stomach contained 10,222, 10,825 and 8,433 tags respectively, as shown in Table 4.1. To permit more direct comparison, each library was normalised to a total of 10,000 tags (at which level, a tag present ten times has an abundance of 0.1%). The normalised, rather than absolute, tag counts are used henceforth.

4.1.1.2 Global analysis of gastric SAGE libraries

Table 4.1 presents a global analysis of the three SAGE libraries. The most common tag overall was CTCCCCCAA, which was identified 781 times (7.8%) in normal stomach and which corresponds to immunoglobulin alpha (IgA). However, on average, only seven genes in each library were expressed at levels of 1% or over. In contrast, the vast majority (97-98%) of transcripts, including the classical house-keeping genes, were present at levels in single figures (below 0.1%). The figures for each library were similar.

4.1.1.3 Global comparison of SAGE libraries from gastric carcinomas and normal stomach

The definition used for the library comparisons was a difference in tag ratio of five-fold or more, with a p-value of 0.01 or less. When tags from the two tumours were pooled and compared with normal stomach, 106 tags were differentially expressed (0.97% of the total genes in the three libraries), of which 47 were higher in the tumours and 59 were higher in normal stomach. When compared individually with normal stomach, 76 tags (1.09%) and 125 tags (1.53%) were differentially expressed in Tumours 1 and 2 respectively, of which 85 tags were up-regulated in either tumour. The two tumours themselves differed by 64 tags (0.75%). These results are in keeping with the histopathological appearances: tumour 1 was better differentiated and of an earlier stage.

4.1.1.4 Global comparison of gastric libraries with other glandular epithelial tissues and mesothelium by digital SAGE

The normal gastric SAGE library was compared digitally with normal breast, colonic, ovarian, pancreatic and prostatic tissue and with normal mesothelium, which lines the peritoneal cavity. 55 genes (0.65% of the total genes in the normal gastric library) were specifically over-expressed in normal stomach.

The pooled gastric adenocarcinomas were compared with the tumour SAGE libraries from: ductal breast carcinoma; ovarian serous adenocarcinoma; colonic, pancreatic and prostatic adenocarcinomas; and malignant mesothelioma of epithelial type. 20 genes (0.23% of the total genes in the pooled gastric tumour libraries) were specifically over-expressed in gastric carcinoma. Initially 22 genes were identified, but two were excluded after they were found in other SAGEmap tumour libraries. Eight gastric-specific genes were common to both tumour and normal comparisons.

Of the 47 tags more highly expressed in the pooled tumours than in normal stomach, only 5 were gastric-specific. In contrast, of the 59 genes more highly expressed in normal stomach compared with the pooled tumours, 33 were gastric-specific.

4.1.1.5 Investigation of individual SAGE tags

Table 4.2 contains the tags differentially expressed between the gastric carcinomas and normal stomach, plus the gastric-specific tags. The 20 most abundant tags in each library are also listed and included many ribosomal and mitochondrial proteins, of which many were more highly expressed in the tumours. Most other transcripts up-regulated in the tumours are expressed in many cell types and will henceforth be termed “widely expressed”. These included growth factors, signal transduction molecules, transcription factors, thymosins and genes involved in protein degradation and invasion. Conversely, most cytoskeletal proteins were down-regulated in the tumours.

Most genes which were more highly expressed in normal stomach than in gastric carcinoma play a role in normal gastric function, and many were gastric-specific compared with other epithelial libraries. The SAGE profiles were thus in part self-validating. Such genes included: gastrin; the anti-bacterial lysozyme; and gastric mucin (MUC5) and trefoil factors pS2 and spasmolytic polypeptide which protect glandular epithelia. Some genes were recently characterised: lipocalin 2 (neutrophil gelatinase-associated lipocalin), prostate stem cell antigen (PSCA) and a new gene called CA11. Others had not been reported in the stomach, such as aquaporin 5. Many differentially or specifically expressed tags corresponded only to uncharacterised cDNAs (i.e. ESTs), or lacked a match in the genetic databases entirely.

A few genes, such as intestinal trefoil factor, were up-regulated in gastric carcinoma, but were not widely expressed; instead, their normal location was the intestine (small and/or large).

4.1.2 Validation of gastric SAGE

4.1.2.1 Northern blotting

To validate and expand the SAGE profiles, selected transcripts were studied in a wider panel of 19 gastro-intestinal tumour and normal tissues and cell lines. Where the genes had been minimally characterised in the stomach, the method of choice was Northern blotting for mRNA, as shown in Figure 4.1. The Northern blots corroborated the SAGE profiles. Gastrin was expressed highly and only in the normal gastric antrum. The new gene CA11 was expressed highly and only in normal stomach, in all three areas: antrum, body (fundus) and cardia. Prostate stem cell antigen (PSCA) was indeed present in normal gastric mucosa and in four of the eight gastric tumours. Lipocalin 2 was present in normal antrum and the two SAGE tumours, although its relative levels differed slightly from those expected.

As predicted, intestinal trefoil factor was identified in seven of the eight gastric tumours, and in normal colon and colonic adenocarcinoma. Of the widely expressed genes, thymosin beta 10 was indeed up-regulated in the gastric and cell lines but was also abundant in the tumour and normal oesophagus and colon. Id1 was highly expressed in the two SAGE tumours, as predicted, but was also present in normal stomach, and was absent from the three cell lines. Prothymosin alpha tended to be more highly expressed in the tumours than in the normal samples but its relative levels in the SAGE samples were rather low. Overall, Northern blotting showed differences in expression levels between tumours, even of the same subtype.

4.1.2.2 Immunohistochemistry

In addition, well-characterised genes were validated by immunohistochemistry (IHC) where antibodies were available. IHC localises the target in tissue sections or cells and identifies the gene's protein product rather than its mRNA transcript. Nevertheless, IHC corroborated the SAGE profiles. Figure 4.2 shows the three SAGE samples stained for gastrin, lysozyme, pS2, spasmolytic polypeptide and MUC5. Table 4.3 lists the IHC results for the full gastro-intestinal panel.

Lysozyme, pS2, spasmolytic polypeptide and MUC5 were highly expressed in normal mucosa from all areas of the stomach. Lysozyme, pS2 and MUC5 were present in all of the distal gastric carcinomas, whether of intestinal or diffuse histological subtype, although their levels varied and were generally lower than in normal tissue. The relative staining by IHC in the three SAGE samples paralleled their SAGE mRNA profiles, with one exception: SAGE had predicted lysozyme to be lacking in Tumour 2, but although some areas were indeed negative by IHC (Figure 4.2), positive staining was present elsewhere (Table 4.3).

The AGS cell line, which is derived from a gastric adenocarcinoma of intestinal type, also expressed lysozyme, pS2 and MUC5, suggesting that it represents a good *in vitro* model. Spasmolytic polypeptide, which was absent or very low in the SAGE tumours, was present in only two tumours in the wider panel, neither of distal, intestinal type. Gastrin was expressed only in the normal antrum, in the G-cells (Figure 4.2). All samples contained cytokeratin 8, as would be expected for simple glandular epithelial tissues, except for normal oesophagus and oesophageal squamous carcinoma, which are composed instead of stratified squamous epithelium. Overall, the IHC staining in the tumours, but not in the normal mucosa, was heterogeneous, both within and between cancers, even those of the same subtype.

4.1.2.3 SAGE tag-to-gene matching

The SAGE tag which represents each mRNA transcript is short, only 13 or 14 bp long (CATG + specific 9 or 10 bp tag), and may correspond to more than one gene. It is therefore important to check that the gene assigned to the tag is correct, ideally through validation studies. Initial searches of on-line cDNA libraries and the background literature are, however, helpful in assessing whether a tag-to-gene match is at least likely, and were used to choose our candidates. Northern blotting and IHC corroborated the SAGE profiles of all but one of our selected genes.

The single problematic tag was ACGCAGGGA, which was very high in the tumours but low in normal stomach (89, 79 and 6 respectively). Although the SAGE program did not generate a match, SAGEmap on-line assigned this tag to the Unigene cluster of heat shock protein 90 alpha (Hsp90a). However, by Northern blotting (Figure

4.1), Hsp90a was present at low levels, with no difference between the SAGE samples. Further investigation showed that, of the three gene sequences in this cluster, one was a chimaera composed partly of Hsp90a and partly of an unknown gene. The tag matched only the latter, which remains uncharacterised. In this instance, the problem with validation arose from the Unigene databases rather than the SAGE profiles, but Unigene remains an extremely useful tool in most cases.

4.2 Discussion

4.2.1 Global analysis and comparison of gastric SAGE libraries

Our data are internally consistent: the gastric SAGE profiles were corroborated by the validation studies. There was, however, heterogeneity within and between tumours. Validation by Northern blotting and immunohistochemistry agreed with the SAGE profiles for 14 out of 15 transcripts. This correlation is at least as good as previous SAGE papers in which between 7% and 50% of differentially expressed tags were not corroborated (Zhang, L. *et al.* 1997; Hibi *et al.* 1998; de Waard *et al.* 1999; Waghray *et al.* 2001). Moreover, although mRNA abundance can be a poor indicator of protein levels (Gygi *et al.* 1999; Pradet-Balade *et al.* 2001), we found that, for a given gene, the relative level of mRNA in different samples correlated well with the relative protein level as gauged by immunohistochemical staining (Figure 4.2). Moreover, while normal tissues from one site varied little, if at all, we found that gene expression varied significantly at both mRNA and protein levels within and between tumours of the same type (Figure 4.1, Figure 4.2, Table 4.2 and Table 4.3). Tumour heterogeneity is a well-recognised phenomenon (Zhang, L. *et al.* 1997; Hibi *et al.* 1998; Lal *et al.* 1999; Perou *et al.* 2000) which explains, at least in part, the reported difficulty in corroborating expression profiles.

Our data are consistent with the literature. The overall distribution of genes in each gastric library is similar to other SAGE libraries. Each of our three SAGE gene expression libraries yielded around 10,000 tags, or transcripts, derived from an average of 4,300 genes (Table 4.1). Only around 7 transcripts were expressed at levels of 1% or more, and over 97% were expressed at levels of 0.1% or less. This pattern of distribution, and the underlying library creation statistics, agree with previous SAGE publications (Zhang, L. *et al.* 1997; Hibi *et al.* 1998; Nacht *et al.* 1999; Velculescu *et al.* 1999; Hough *et al.* 2000; Waghray *et al.* 2001).

Only 1% of transcripts were differentially expressed between gastric carcinoma and normal stomach. This figure for the pooled gastric tumour libraries is similar to previous SAGE papers comparing colonic, pancreatic, breast, ovarian, prostatic and lung carcinomas with normal tissues, although the criteria for differential expression varied (Zhang, L. *et al.* 1997; Hibi *et al.* 1998; Nacht *et al.* 1999; Velculescu *et al.* 1999; Hough *et al.* 2000; Waghray *et al.* 2001). When compared individually with normal stomach, the moderately differentiated gastric carcinoma contained more differentially expressed tags (T2: 125 tags) than the well-differentiated tumour (T1: 76 tags), thus the molecular anatomy of the tumours is in keeping with their morphology. To our knowledge, this correlation between the tumours' differentiation state and global mRNA profile has not previously been described, although it makes biological sense.

The correlate of the fact that only 1% of genes were differentially expressed between normal and malignant tissues is that the levels of 99% of transcripts were similar. These genes would presumably be unlikely to make good diagnostic or therapeutic targets, unless of course the corresponding protein differed in its abundance, stability or functional state.

Up to 0.65% of transcripts were gastric-specific. When normal stomach was compared with SAGE libraries from normal glandular epithelia (breast, colon, ovary, pancreas and prostate) and mesothelium, 55 gastric-specific genes (0.65%) emerged. In their paper on the "human transcriptome", Velculescu *et al* re-analysed the SAGE data available in 1999 and suggested that tissue-specific genes in normal samples vary in abundance from 0.09% (keratinocytes) to 1.76% (colon) (Velculescu *et al.* 1999). Our figure is in keeping with this range.

Genes specifically over-expressed in gastric adenocarcinoma compared with adenocarcinomas of the breast, colon, ovary, pancreas and prostate and malignant mesothelioma were fewer in number: 20 genes (0.24%). Comparable data for tissue-specific genes in tumours, rather than in normal tissue, were not included in the "human transcriptome" SAGE paper and are difficult to find. However, this topic was studied using the data-mining tool called Digital Differential Display (DDD) on NCBI's web-site (Wheeler *et al.* 2001). ESTs from adenocarcinomas of the breast, colon, lung, ovary, pancreas and prostate were analysed for genes specific for tumour type, by comparison with pooled normal and tumour libraries (Scheurle *et al.* 2000). Over 80 transcripts were specifically up-regulated in the six tumour types, which averages to around 16 genes per tissue (Scheurle *et al.* 2000). Again, this estimate is consistent with our figure.

That there were fewer tissue-specific genes in gastric carcinoma than in normal stomach, and in adenocarcinomas overall compared to their corresponding normal glandular epithelial tissues, makes biological sense and is in keeping with tumours in general reverting to a less specialised, that is, less differentiated state.

4.2.2 Investigation of individual SAGE tags in gastric libraries

Ribosomal and mitochondrial genes were amongst the most abundant transcripts in all three libraries and were generally up-regulated in the tumours. This agrees with previous general expression profiles and comparisons of normal and malignant tissues by

SAGE and microarrays (Zhang, L. *et al.* 1997; Hibi *et al.* 1998; de Waard *et al.* 1999; Velculescu *et al.* 1999; Perou *et al.* 2000), and with prior EST data (Adams *et al.* 1995). The data are also consistent with previous global comparisons of normal stomach and gastric carcinoma, including four differential display (DD) (Ebert *et al.* 2000; Jung *et al.* 2000; Yoshikawa *et al.* 2000; Wang *et al.* 2001) and one membrane array (El-Rifai *et al.* 2001) studies. In the DD study which identified the most differentially expressed transcripts, six out of 13 genes up-regulated in gastric carcinoma were ribosomal or mitochondrial (Jung *et al.* 2000). The single gene identified in another DD comparison of non-metastatic and metastatic gastric cancer cell lines was also mitochondrial (Salesiotis *et al.* 1995). That the most abundant transcripts overall were involved in protein synthesis and energy production, and that these activities were increased in malignancy, in which cells are less differentiated and abnormally proliferative, also makes biological sense.

There were marked differences in the abundances of transcripts for ribosomal proteins, which are thought to be present in near equal amounts in each ribosomal particle. Similar results have been seen in most other SAGE analyses (Yamamoto *et al.* 2001). The reasons are unclear but presumably might reflect different rates of turnover of each ribosomal subunit protein.

Most other transcripts up-regulated in the tumours were "widely expressed" and included transcription factors and genes involved in cell signalling and protein degradation. Only 5 of the 47 tags up-regulated in the pooled tumours were gastric-specific, compared with 33 of 59 tags more highly expressed in normal stomach. The up-regulation in tumours of widely expressed, or constitutive, transcripts is again in keeping with cellular de-differentiation during malignant transformation. The genes included transcription factors such as c-myc binding protein, zinc finger protein homologous to murine Zfp-36, and Id1.

Id1. Id1 was highly expressed in the gastric carcinomas, but was also present at lower levels in normal stomach. Id1 is a member of the ubiquitously expressed family of ID (inhibitor of differentiation) proteins which inhibit the basic helix-loop-helix transcription factors (Norton 2000). ID proteins are implicated in the regulation of tumour growth, angiogenesis, invasion and metastasis and increased Id1 levels correlate with a more aggressive phenotype in breast carcinoma (Norton 2000). Up-regulation of Id1 in gastric carcinoma could thus have been predicted but its absence from the gastric cancer cell lines is surprising.

Thymosins. Thymosin beta 10 and prothymosin alpha were also up-regulated in the gastric tumours. Thymosins are small proteins which were originally isolated from the thymus and were considered to have an immune function (Huff *et al.* 2001). Prothymosin alpha is now thought to play a role in cell proliferation, although its exact function remains unclear; its up-regulation in gastric cancer has been previously reported (Mitani *et al.* 2000). The beta thymosins bind monomeric (globular) actin and thymosin beta 10 is known to be up-regulated in carcinomas compared with normal tissues, although the stomach has not previously been studied (Huff *et al.* 2001). Thymosins are interesting targets, against which drugs are already in development (Huff *et al.* 2001).

Intracellular trafficking and protein degradation. Coatomer and proteasome components were more highly expressed in the tumours. Vesicles bearing coatomer proteins cycle between the endoplasmic reticulum and Golgi complex (Wu *et al.* 2000), and contain newly synthesised proteins awaiting post-translational modification, so that coatomer up-regulation is in keeping with enhanced protein synthesis in tumours. The proteasome is the main cellular proteolytic machinery and could contribute to malignant transformation by altered degradation of, for example, APC (Schwartz *et al.* 1999).

Cell signalling. Other genes with well-recognised roles in carcinogenesis were also up-regulated in the tumours: collagenase I, which facilitates invasion; growth factors (hepatocyte growth factor) and their ligands (fibroblast growth factor receptor) (see Section 1.5.4); and signal transducers including guanylate kinase and putative serine/threonine protein kinases. The latter are interesting since they and the related tyrosine kinases are often over-expressed in cancers, may be relatively tissue-specific, and have proven to make excellent therapeutic targets, with STI571 for example (see Section 1.5.6) (Blume-Jensen *et al.* 2001).

Overall profile of gastric carcinoma. The genes more highly expressed in gastric tumours are mostly widely expressed, and many of them interact: for example, serine/threonine protein kinases are linked both to the cytoskeleton and to cascades which alter transcription down-stream, so that identification of the few critical underlying molecular events may be challenging (Hanahan *et al.* 2000). Our profiles are consistent with the general cancer literature (Hanahan *et al.* 2000) and with the cDNA membrane array study of gastric tumours, in which up-regulated transcripts included growth factors and genes involved in the cell cycle, adhesion and invasion (El-Rifai *et al.* 2001). This supports the existence of common molecular targets in cancers for diagnosis and therapy.

Unlike other widely expressed genes, most cytoskeletal proteins were highly expressed in normal stomach and were down-regulated in the tumours. Keratin 8 is an intermediate filament specific to simple glandular epithelia (Green *et al.* 2000). Intermediate filaments are anchored at the cell membrane through junctions, including desmosomes, of which desmoplakin and cadherins are major components (Green *et al.* 2000). Alpha actinin, profilin, cofilin and gelsolin contribute to the actin microfilament cytoskeleton (Janke *et al.* 2000). All of these genes were abundant in normal stomach and down-regulated in the tumours. This altered expression pattern is shared with other carcinomas and indeed functional assays suggest that many cytoskeletal proteins act as tumour suppressor genes (Janke *et al.* 2000), as is the case with E-cadherin (see Section 1.5.5). Where beta thymosins fit into this scheme is unclear: they are also actin-binding proteins but are up-regulated in tumours, and may enhance cell motility (Huff *et al.* 2001).

Most genes which were more highly expressed in normal stomach play a role in gastric function. Many were also gastric-specific. It is worth remembering the functions of the stomach (see Section 1.2.1). It acts as a reservoir for food and mechanically churns and mixes it with gastric juice containing: hydrochloric acid which is a sterilising agent, denatures proteins and activates digestive enzymes; the protease pepsin; intrinsic factor; and gastric lipase. The stomach is not a major absorptive site, but water, ions, short chain fatty acids and alcohol are absorbed. The stomach must also protect its mucosal lining from proteolytic and acid attack, and its activity is co-ordinated with the rest of the gut through neuro-hormonal mechanisms.

Gastrin. Gastrin is one such gut hormone. We found gastrin only in normal antral mucosa, and it was gastric-specific. This is as expected: gastrin is secreted by antral G-cells in response to food entering the stomach and it stimulates the gastric body mucosa to secrete acid and pepsinogen A (see Section 1.2.3). Gastrin is only occasionally identifiable in gastric adenocarcinomas although it is a major product of gastric carcinoid (neuro-endocrine) tumours (Berner *et al.* 1991).

Mucosal immune defence. Defence against micro-organisms is an important role of the gut (Owen 1986). Immunoglobulins, especially IgA splice variants, and the anti-bacterial proteins lysozyme and lipocalin 2, were abundant in normal stomach (the SAGE sample had a chronic gastritis, without dysplasia: see Section 1.3.1) and were mostly down-regulated in the tumours and gastric-specific. IgA and lysozyme are well-characterised and known to be highly expressed in gastritis and in carcinomas of intestinal type, especially when well differentiated as in Tumour 1 (Isaacson 1982). Lipocalins have been identified more recently and bind small lipophilic molecules, including bacterial-derived lipopolysaccharides (Friedl *et al.* 1999). Lipocalin 2 has been reported in normal glandular epithelia (Friedl *et al.* 1999) but not, until now, in cancers.

Mucins and trefoil peptides. The mucin MUC5 and trefoil factors pS2 (trefoil factor 1 or TFF1) and human spasmolytic polypeptide (TFF2) were highly expressed in normal stomach, were down-regulated in the tumours, especially in the less well differentiated Tumour 2, and were also gastric-specific. However, although pS2 emerges as specific to stomach compared with other normal tissues, it is also over-expressed in breast and other adenocarcinomas (Wong *et al.* 1999). Mucins are high molecular weight glycoproteins, and trefoil factors (TFF) are small peptides which are resistant to acid, enzymes and heat (Wong *et al.* 1999; Machado *et al.* 2000). Both are synthesised by glandular epithelia, where they are normally co-expressed in a site-specific manner. Mucins and TFFs act synergistically to protect and repair mucosal surfaces. In normal stomach, MUC5 and pS2 are co-localised in the superficial epithelium while spasmolytic polypeptide is present in the antral glands (with MUC6) (Machado *et al.* 2000), as seen in Figure 4.2. In a previous study of 96 gastric carcinomas, around 60% of tumours expressed pS2 and MUC5, but only 10% contained spasmolytic polypeptide (Machado *et al.* 2000), in keeping with our findings. Moreover, in a knockout mouse model, homozygous pS2-null mice develop hyperplasia, dysplasia and carcinoma of the gastric antrum, which suggests that pS2 is important in normal gastric function and may act as a gastric-specific tumour suppressor gene (Lefebvre *et al.* 1996). It is possible that some of the less well characterised genes identified in this study may play a similar role.

Pepsinogens. All three specimens contained pepsinogens, which are the inactive precursors of the gastric protease pepsin (Owen 1986). Perhaps the best-recognised form is pepsinogen A, which is normally secreted by the chief cells in the gastric body, whereas pepsinogen C is produced in all areas of the stomach (Konishi *et al.* 1995). Pepsinogens of either type are present in 20% of gastric carcinomas (Konishi *et al.* 1995). We found pepsinogen A only in Tumour 2 whereas pepsinogen C was present in normal gastric antrum and in Tumour 1. Both pepsinogens were gastric-specific.

Prostate stem cell antigen (PSCA). PSCA was expressed in normal stomach and down-regulated in gastric carcinomas. This agrees with a recent paper characterising PSCA, which was first identified as a protein over-expressed in prostate cancer (Bahrenberg *et al.* 2000). (Note that PSCA is a different gene from prostate-specific antigen (PSA)). Its function is unknown but a role in cell adhesion has been suggested (Bahrenberg *et al.* 2000).

Transport of water and ions. Aquaporin 5 was expressed in all three SAGE samples but more highly in normal antrum and Tumour 1. Fluid transport is a major function of the gut and over nine litres are absorbed or secreted across its epithelia daily (Ma *et al.* 1999). The aquaporins are integral membrane proteins which act as water channels, and at least seven family members are found in different parts of the gut (Ma *et al.* 1999). Aquaporin 5 has not previously been identified in the stomach but is present in the salivary glands (Ma *et al.* 1999). Proteolipid protein 2 (colonic epithelium-enriched A4 protein) was also present in normal stomach and down-regulated in the tumours. In the rat, it is expressed in the intestine (Breitwieser *et al.* 1997). Proteolipid protein 2 shows features of an ion channel (Breitwieser *et al.* 1997), and although not previously described in the stomach or in humans, is abundant in gastric EST libraries in Unigene.

Most transcripts more highly expressed in normal stomach (Table 4.2) were thus involved in gastric function. In some cases, the genes were well characterised, and for others their role could be inferred from existing data on gastro-intestinal physiology. However, many new genes were identified as well as known genes with unexpected expression patterns.

New gastric gene CA11. One tag was expressed at very high levels in normal stomach, was absent from the tumours and was gastric-specific. Initially this tag matched only ESTs, which were used to create a cDNA probe for Northern blotting which confirmed the SAGE results. Simultaneously the gene was identified through DD by a Japanese group who named it CA11 (which does not indicate carbonic anhydrase) and reported the same expression pattern (Yoshikawa *et al.* 2000). CA11 is one of the most abundant transcripts in normal stomach, at levels similar to mucin, which suggests that it has an important role in normal gastric function. CA11 is a good example of the power of mRNA profiling in identifying novel genes, and its further characterisation is described in Chapter 5.

One gene expressed in the gastric carcinomas was thought until recently to be specific to ovarian and mesothelial tissues. Mesothelin was expressed in gastric carcinomas but was absent from normal stomach. Mesothelin is a glycoprotein which may function in cell adhesion (Chang *et al.* 1996). It is present in normal mesothelium and malignant mesothelioma (Chang *et al.* 1996), and is also highly expressed in ovarian carcinomas compared with non-transformed ovarian epithelium (Hough *et al.* 2000) and with other adenocarcinomas (Scheurle *et al.* 2000). Mesothelin has thus been proposed as an ovarian/mesothelial-specific marker, but this is contradicted by this study and others: mesothelin has recently been identified in gastric cancer tissues (El-Rifai *et al.* 2001) and cell lines (Hippo *et al.* 2001), in which its expression is associated with peritoneal metastasis (Hippo *et al.* 2001). This is not surprising given the normal location of mesothelin, which may represent a candidate therapeutic target.

Some genes more highly expressed in the gastric carcinomas were in keeping with acquisition of an intestinal phenotype. Intestinal trefoil factor was absent from normal stomach but was up-regulated in the gastric carcinomas; this, the third TFF, is normally expressed in the small intestine and colon (Wong *et al.* 1999). This expression pattern was shared with other genes. Sulfotransferase 1A1 is an enzyme which detoxifies xenobiotics and endogenous compounds (Harris *et al.* 2000), and is normally expressed mainly in the intestines. Butyrate response factor 2 is normally expressed in the colon where the short-chain fatty acid butyrate is produced by bacterial fermentation of luminal carbohydrates and acts as the main mucosal energy source (Gibson *et al.* 1999). The up-regulation in gastric carcinoma of transcripts normally associated with the distal gut, especially the colon, presumably reflects the histological intestinal tumour phenotype (see Section 1.4.5) (Laurén 1965; Fuchs *et al.* 1995). Once again, the molecular anatomy of the tumours is in keeping with their morphology.

Some tissue-specific genes associated with normal gastric antral mucosa are highly expressed in pancreatic adenocarcinomas. The three genes highlighted in a recent SAGE study as being up-regulated in pancreatic adenocarcinoma, compared with normal pancreatic ductal cells, were spasmolytic polypeptide, lipocalin 2 and PSCA (Argani *et al.* 2001). The DDD study additionally identified MUC5 and MUC1 as being over-expressed in pancreatic adenocarcinoma (Scheurle *et al.* 2000). These genes have been proposed as new markers of pancreatic cancer. In fact, all five were highly expressed in normal stomach (and were down-regulated in gastric carcinoma). Pancreatic adenocarcinomas have previously been shown to express antigens normally found in gastric antral mucosa, such as MUC5 and pepsinogen C (Sessa *et al.* 1990), and a gastric antral phenotype commonly develops in other mucinous tumours, of the ovary (Tenti *et al.* 1992) and endocervix (Tenti *et al.* 1994).

Conversely, gastric mucosa may show metaplasia to the acinar (exocrine) tissue of the pancreas (Doglioni *et al.* 1993); such so-called pancreatic metaplasia, like the more common intestinal metaplasia (see Section 1.3.1) tends to occur in association with chronic gastritis. Interestingly, *H. pylori* infection has now been shown to be associated with an increased risk of pancreatic adenocarcinoma (as well as gastric carcinoma, obviously) (Stolzenberg-Solomon *et al.* 2001). The underlying mechanism is unclear but could be related to the increased gastrin levels frequently found in atrophic gastritis: as well as its stimulation of gastric acid secretion and motility, gastrin is trophic to most of the gut (Stolzenberg-Solomon *et al.* 2001).

The reasons why gastric carcinomas should frequently assume an intestinal phenotype, and why pancreatic carcinomas show features of normal gastric antral mucosa, are unclear but may reflect aberrant expression of the developmental homeobox genes (Beck *et al.* 2000). For distal gastric carcinomas, this presumably would act at an early stage in Correa's carcinogenic pathway, from intestinal metaplasia onwards. In any case, these phenomena emphasise the need to examine many different cells and tissues when evaluating potential markers, preferably including techniques which enable localisation of the target gene.

Some tags which were highly, differentially or specifically expressed in the

gastric SAGE libraries correspond to multiple genes. Many others lacked a match entirely. Both issues can be addressed through SAGEmap, Unigene and Genbank (Lal *et al.* 1999; Wheeler *et al.* 2001). For multiple matches, certain gene assignments can be judged more likely than others, based on their function and expression elsewhere, before selecting candidates for validation studies. Tags lacking matches to well characterised genes often correspond to ESTs, and new matches will emerge from the ever-expanding genetic databases (Lal *et al.* 1999; Wheeler *et al.* 2001). A new method which generates longer, more specific cDNA fragments from the short SAGE tags should also be useful (Chen *et al.* 2000).

Conclusion. These are the first global profiles of gene expression in gastric carcinoma and normal stomach created using serial analysis of gene expression (SAGE). Two libraries of gastric adenocarcinoma of distal, intestinal type, and one library of normal gastric antrum have been produced. Numerous transcripts have been identified which are: highly expressed; differentially expressed between normal and tumour stomach; or gastric-specific by comparison with normal and tumour breast, colon, ovary, pancreas, prostate and mesothelium. Selected genes have been validated in a wider panel of 19 gastro-intestinal tissues by Northern blotting and immunohistochemistry.

The overall statistics of the three SAGE libraries agreed with previous publications, with 1% of genes being differentially expressed between gastric carcinoma and normal stomach. The SAGE profiles were corroborated by the validation studies. The most abundant transcripts included ribosomal and mitochondrial proteins, of which most were up-regulated in the tumours, as were other widely expressed genes including transcription factors, growth factors, and genes involved in signal transduction, protein turnover and cell invasion. This pattern is similar to other cancers, which supports the existence of common molecular targets for diagnosis and therapy. Most genes which were abundant or more highly expressed in normal stomach play a role in normal gastric function, including gastrin, lysozyme, mucins, trefoil factors and pepsinogens, of which some (up to 0.65%) were gastric-specific by comparison with other normal glandular tissues.

The molecular anatomy of the tumours correlated with their morphology. The expression profile of the well differentiated gastric carcinoma more closely resembled normal stomach than did the moderately differentiated tumour. Some genes up-regulated in the gastric carcinomas indicated the acquisition of an intestinal phenotype, which has long been recognised histologically. Some genes associated with normal gastric antrum were also abundant in, and had previously been proposed as specific markers of, pancreatic carcinoma. These expression profiles, unexpected in some cases, together with tumour heterogeneity, emphasise the need for candidates to be tested, and ideally localised, in a wide range of normal and diseased cells and tissues.

New candidate genes have been identified. Some transcripts had previously been characterised minimally or not at all in the stomach. Many SAGE tags with interesting expression patterns, some tumour-associated, lacked matching genes and await further characterisation and functional studies. These molecular portraits increase our knowledge about the genes involved in normal gastric function and in malignant change in the stomach, and provide a catalogue of candidates from which to develop markers for better diagnosis and therapy of gastric carcinoma.

In the next Chapter, 5, I go on to characterise in detail a novel gene which was highly expressed in normal stomach.

Figure 4.1: Northern blotting of selected genes identified by SAGE

RNA was isolated from 19 gastro-intestinal tumor and normal tissues and cell lines, as indicated along the top row, of which further details are listed in Table 4.3.

The Northern blots corroborated the SAGE profiles. Gastrin was expressed highly and only in the normal gastric antrum. The new gene CA11 was expressed highly and only in normal stomach, in all three areas: antrum, body and cardia (see Section 1.2.1). Prostate stem cell antigen (PSCA) was indeed present in normal gastric mucosa and in four of the eight gastric tumours. Lipocalin 2 was present in normal antrum and the two SAGE tumours, although its relative levels differed slightly from those expected.

As predicted, intestinal trefoil factor was identified in seven of the eight gastric tumours, and in normal colon and colonic adenocarcinoma. Of the widely expressed genes, thymosin beta 10 was indeed up-regulated in the gastric and cell lines but was also abundant in the tumour and normal oesophagus and colon. Id1 was highly expressed in the two SAGE tumours, as predicted, but was also present in normal stomach, and was absent from the three cell lines. Prothymosin alpha tended to be more highly expressed in the tumours than in the normal samples but its relative levels in the SAGE samples were rather low. Overall, Northern blotting showed differences in expression levels between tumours, even of the same subtype.

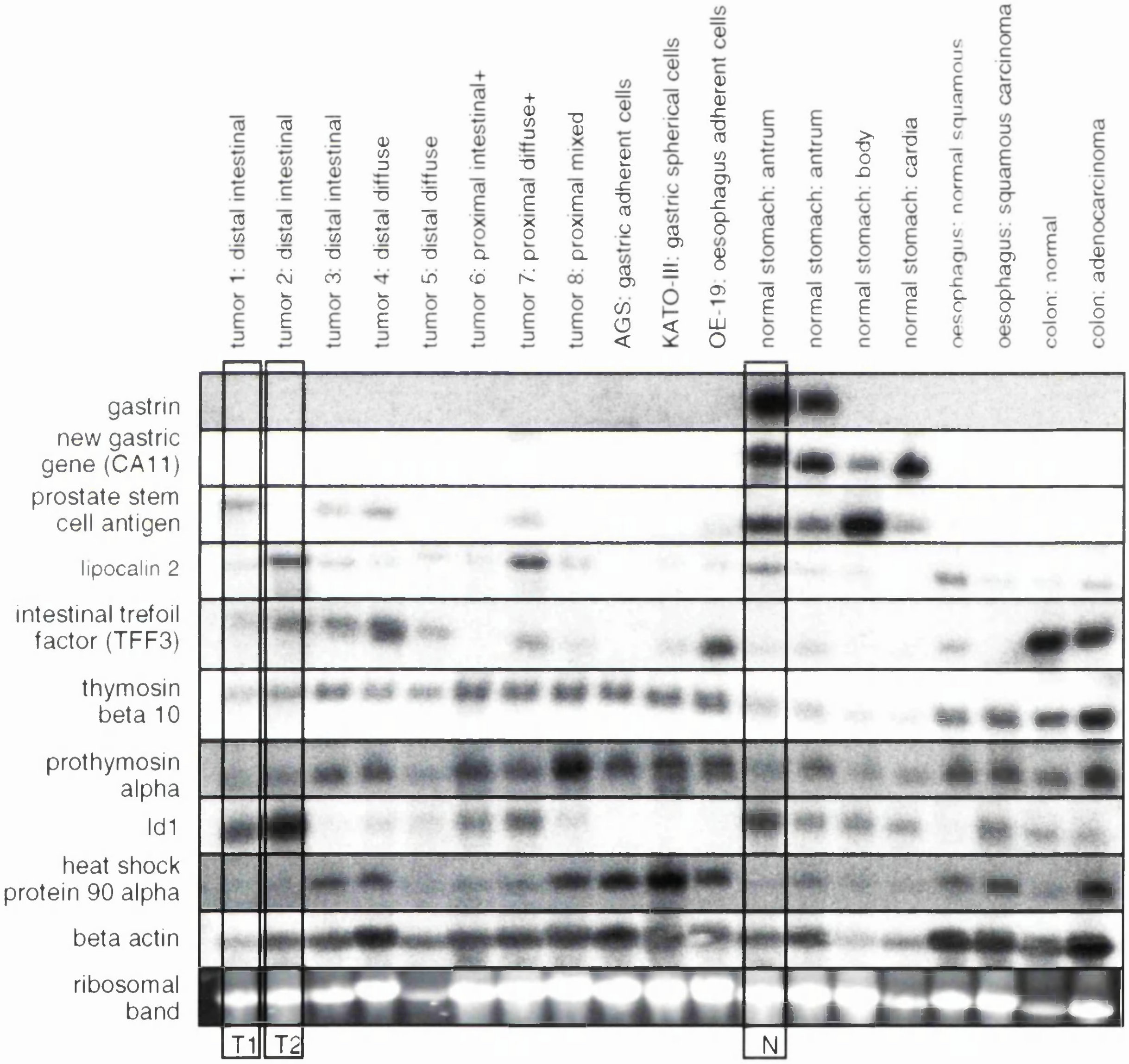

tumor 1: distal intestinal
tumor 2: distal intestinal
tumor 3: distal intestinal
tumor 4: distal diffuse
tumor 5: distal diffuse
tumor 6: proximal intestinal+
tumor 7: proximal diffuse+
tumor 8: proximal mixed
AGS: gastric adherent cells
KATO-III: gastric spherical cells
OE-19: oesophagus adherent cells
normal stomach: antrum
normal stomach: antrum
normal stomach: body
normal stomach: cardia
oesophagus: normal squamous
oesophagus: squamous carcinoma
colon: normal
colon: adenocarcinoma
gastrin
new gastric gene (CA11)
prostate stem cell antigen
lipocalin 2
intestinal trefoil factor (TFF3)
thymosin beta 10
prothymosin alpha
Id1
heat shock protein 90 alpha
beta actin
ribosomal band
T1
T2
N

Figure 4.2: Immunohistochemistry for selected genes identified by SAGE

Immunohistochemistry (IHC) was used to validate the better characterised genes where antibodies were available. IHC localises the target in tissue sections or cells and identifies the gene's protein product rather than its mRNA transcript. Nevertheless, IHC corroborated the SAGE profiles.

A wide range of 19 gastro-intestinal tumour and normal tissues and cell lines was stained. These photomicrographs depict representative areas from the two distal, intestinal tumours (T1 and T2) and normal stomach (antrum) (N) which were subjected to SAGE. The three SAGE samples were stained for gastrin, lysozyme, pS2 (TFF1), spasmolytic polypeptide (TFF2) and MUC5. Spasm polypep stands for human spasmolytic polypeptide (TFF2). The magnification is the same throughout. The number in the bottom left-hand corner is the normalised SAGE tag count for comparison. Table 4.3 lists the IHC results for the full panel of 19 gastro-intestinal samples.

Lysozyme, pS2, spasmolytic polypeptide and MUC5 were highly expressed in normal mucosa from all areas of the stomach. pS2 (TFF1) and MUC5 are co-expressed in the superficial/foveolar epithelium, whereas spasmolytic polypeptide (TFF2) is located at the base of the glands, with fainter staining.

Lysozyme, pS2 and MUC5 were present in all of the distal gastric carcinomas, whether of intestinal or diffuse histological subtype, although their levels varied and were generally lower than in normal tissue. The relative staining by IHC in the three SAGE samples paralleled their SAGE mRNA profiles, with one exception. SAGE had predicted lysozyme to be lacking in Tumour 2, but although some areas were indeed negative by IHC, positive staining was present elsewhere: this variability is evident even in this single photomicrograph. Spasmolytic polypeptide, which was absent or very low in the SAGE tumours, was present in only two tumours in the wider panel, neither of distal, intestinal type. Overall, the IHC staining in the tumours, but not in the normal mucosa, was heterogeneous, both within and between cancers, even those of the same subtype.

Gastrin was expressed only in the normal antrum, in the G-cells in the lower parts of the glands, as seen here.

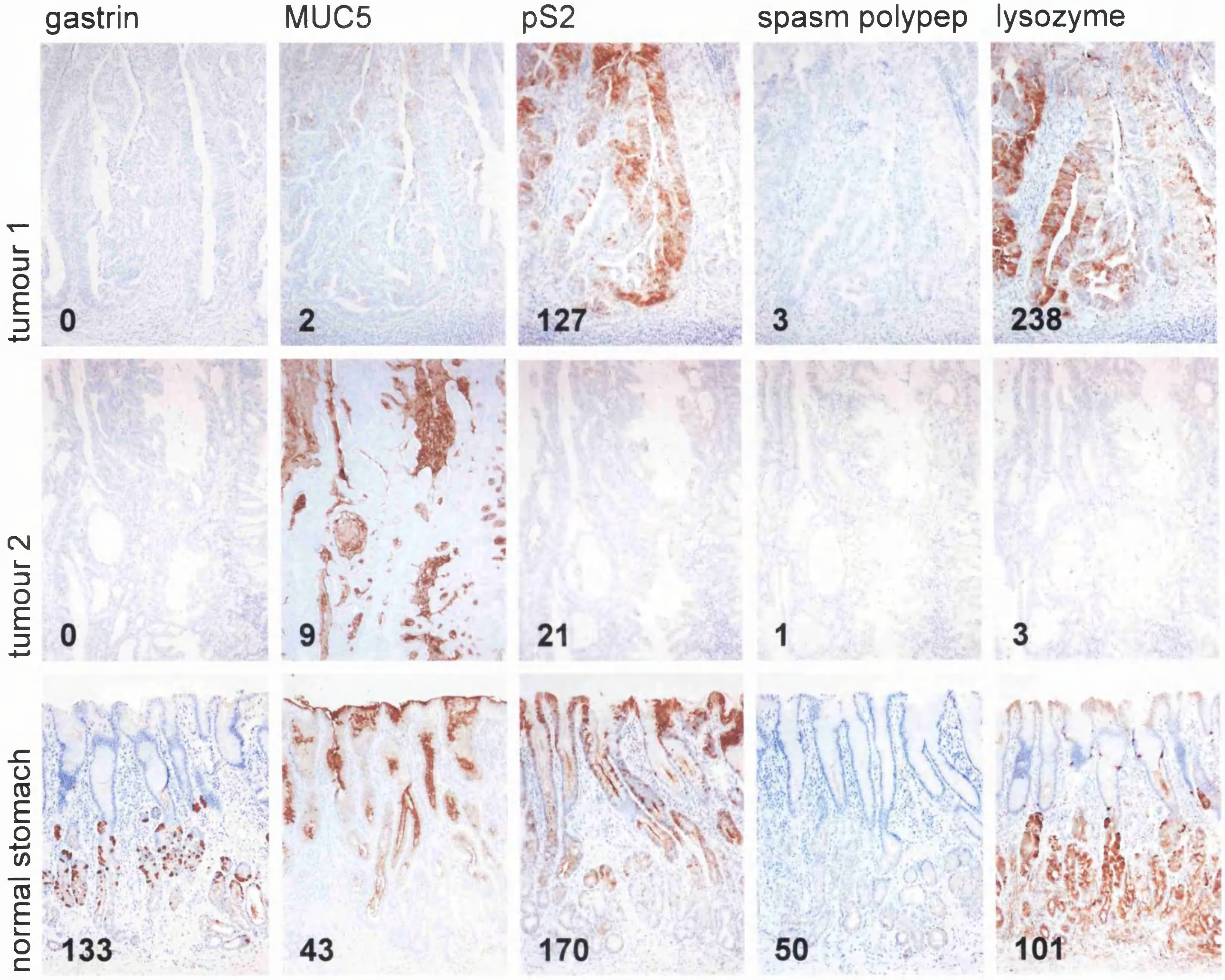

gastrin
MUC5
pS2
spasm polypep
lysozyme
tumour 1
0
2
127
3
238
tumour 2
0
9
21
1
3
normal stomach
133
43
170
50
101

Table 4.1: Summary of generation and overall analysis of SAGE tag libraries

Three SAGE tag libraries were generated. Two were from gastric adenocarcinomas, both of distal, intestinal type (T1 and T2, which simply stand for Tumours 1 and 2, and do not refer to the stage of the tumours: for further details, see Table 2.1). The third was from normal gastric antral (distal) mucosa (N), which had a chronic gastritis, without dysplasia.

Sequences indicated the number of cloned SAGE concatemer inserts which were sequenced. From these concatemers are derived the SAGE tags, which average around 10,000 for each library. The % abundance of tags refers to those present at or over 1%, 0.5% and 0.1%: in a library normalised to a total of 10,000 tags, these figures translate to an abundance of 100, 50 and 10 tags or over respectively.

The absolute number of unique tags obtained is also given. Because of inevitable sequencing errors, this is usually regarded as a slight over-estimate of the true number of different genes expressed and a correction which removes around 7% of tags is sometimes applied (Velculescu *et al.* 1995; Zhang, L *et al.* 1997; Velculescu *et al.* 2000).

Library	Sequences	Tags	Tags at ≥1%	Tags at ≥0.5%	Tags at ≥0.1%	≥ Two tags	Unique tags
Tumour 1 (T1)	825	10222	7	21	124	1008	4284
Tumour 2 (T2)	690	10825	6	16	115	1008	5350
Normal (N)	987	8433	7	19	107	822	3671

Table 4.2: Tags which were highly expressed, differentially expressed between gastric carcinomas and normal tissue, specifically expressed in the stomach, or otherwise of interest

The tags listed are 9 bp rather than 10 bp in length, although 10 bp tags are used in most SAGE publications. This is for technical reasons, as described in Section 3.1.4.

In the Gene Match column, where a tag is assigned to a gene, it clearly matches a well-characterised mRNA transcript or gene. Matching was performed using either the SAGE software with Genbank or SAGEmap, as described in Sections 2.2.3.2 and 3.1.3. An *asterisk indicates that gene expression has been validated by Northern blotting or immunohistochemistry. The genes have been sub-divided partly by function, for which further investigation will be enabled by the new Gene Ontology databases (Ashburner *et al.* 2000). The section "Multiple matches" contains tags for which there are numerous possible matching genes. The section "No matches" contains tags for which a matching well-characterised gene is lacking. For these tags, any matching ESTs have been listed, along with the tag abundance in other SAGE libraries. In some cases, the tag matches ESTs but not well-characterised cDNAs in a Unigene cluster.

In the next three columns, the two tumor and the normal samples are represented by T1, T2 and N respectively. The absolute tag counts are normalised to 10,000 total tags per library to permit direct comparison. At this scale, the classical housekeeping genes are at low single figure levels (0.01-0.1%). Genes expressed highly, differentially or (tissue-) specifically in the tumors are shaded in mid-grey whereas the equivalents in normal stomach are shaded in light grey. Genes which are highly or (tissue-)specifically expressed in both tumour and normal libraries but which are not differentially expressed are shaded in black. Tissue-specific expression is by digital comparison with other normal and tumorous glandular epithelial tissues, such as colon and breast, and mesothelium (see Section 4.1.1.4).

The 20 most abundant ("high") tags in each of the tumour and normal libraries are then listed. Ts indicates both tumours.

Tags with differential expression between the tumours and normal stomach was defined as a difference of five-fold or more combined with a p-value of 0.01 or less. "Pool" indicates that the genes were differentially expressed in a comparison using the pooled tumours. "T either" indicates that the tag was differentially expressed only in an individual comparison between one or other tumour and normal gastric antrum.

Tags with tissue-specific differential gene expression were defined in the same way, as a difference of five-fold or more combined with a p-value of 0.01 or less.

					20 most abundant tags		Differential expression		Tissue-specific expression	
Tag	Gene match	T1	T2	N	T	N	T	N	T	N
	Ribosomal proteins and related genes involved in translation									
GTGAAGGCA	ribosomal protein S3a	26	55	8	T2 high		T either			
CCTTCGAGA	ribosomal protein S5	1	10	1			T either			
GCCGTGTCC	ribosomal protein S6	17	39	7	T2 high		T either			
TCGTCTTTA	ribosomal protein S7	2	7	0			T either			
GCCGAGGAA	ribosomal protein S12	61	92	56	Ts high	N high				
GTGTTGCAC	ribosomal protein S13	2	14	0			T pool			
CCGTCCAAG	ribosomal protein S16	31	53	15	T2 high					
GCTCCGAGC	ribosomal protein S16 (possible second tag)	1	7	0			T either			
TGGTGTTGA	ribosomal protein S18	44	159	8	T2 high		T pool			
GCTTTTAAG	ribosomal protein S20	6	13	2			T either			
CTGTTGGTG	ribosomal protein S23	7	7	0			T pool			
TAAGGAGCT	ribosomal protein S26	6	10	0			T pool			
CTCAGACAG	ribosomal protein S27	4	11	0			T pool			
CGCCGGAAC	ribosomal protein L4	1	5	23				N pool		
CTGCTATAC	ribosomal protein L5	3	8	0			T either			
TACAAGAGG	ribosomal protein L6	1	7	0			T either			
ATCAAGGGT	ribosomal protein L9	5	18	4			T either			
GTGTTAACC	ribosomal protein L10	3	15	0			T pool			
ACATCATCG	ribosomal protein L12	1	13	2			T either			
CCCGTCCGG	ribosomal protein L13	15	24	55		N high				
GGAGTGGAC	ribosomal protein L18	0	10	0			T either			
AAGGTGGAG	ribosomal protein L18a	32	75	42	T2 high					
AAGGTCGAG	ribosomal protein L24	3	12	0			T pool			
CCCATCCGA	ribosomal protein L26	4	30	4			T either			
GGGCTGGGG	ribosomal protein L29 or others	19	16	95		N high		N pool		
GGGTTGGGG	ribosomal protein L29 or others	0	0	7				N pool		
AAGGAGATG	ribosomal protein L31	33	36	4			T pool			
CGCCGCCGG	ribosomal protein L35	29	45	58	T2 high	N high				
GTTCGTGCC	ribosomal protein L35a	10	10	1			T pool			
TTACCATAT	ribosomal protein L39	0	8	0			T either			
TTGGTCCTC	ribosomal protein L41	64	72	38	Ts high					
TTGGCCCTC	homologue to yeast ribosomal protein L41	3	0	15				N pool		N spec
TGTGTTGAG	elongation factor 1 alpha	83	117	64	Ts high	N high				
CGCCGCGGT	translation initiation factor eIF-3 p110 subunit	2	2	26				N pool		
GTGACAGAA	eukaryotic translation initiation factor 4A, isoform 1	5	5	0						
CCCCCTCCG	small nuclear ribonucleoprotein (snRNP) B1	2	1	0						
GCCCAGCTG	eukaryotic translation elongation factor 1 delta	1	0	7				N pool		
GGCCCTGAG	RNA polymerase II subunit	6	4	9						
	Mitochondrial proteins									
CCCATCGTC	mitochondrial cytochrome oxidase II	89	146	74	Ts high	N high				
TGATTTCAC	mitochondrial cytochrome oxidase III	9	35	0			T pool			
CACTACTCA	mitochondrial cytochrome b	0	37	0			T pool			
CACCTAATT	mitochondrial ATPase 6/8	68	89	23	Ts high					
ACTAACACC	mitochondrial NADH dehydrogenase II	13	59	0	T2 high		T pool			
TTCATACAC	mitochondrial DNA loop attachment sequence	23	48	8	T2 high		T either			
TCGAAGCCC	mitochondrial DNA loop attachment sequence	6	23	0			T pool			
GGGGACTGA	low molecular mass ubiquinone-binding protein d50369	2	12	0			T either			
TTGGAGATC	NADH:ubiquinone oxidoreductase MLRQ subunit	9	0	0			T pool			
ACCCTTGGC	tag matches mitochondrial sequence	0	67	5	T2 high		T pool			
AGACCCACA	tag matches mitochondrial sequence	0	12	0			T pool			
ATTTGAGAA	tag matches mitochondrial sequence	1	15	1			T either			
CAAGCATCC	tag matches mitochondrial sequence	54	37	4			T pool			
CTCATAAGG	tag matches mitochondrial sequence	20	4	0			T pool			
	Transcription factors and other nuclear proteins									
CAGCAGAAG	small EDRK-rich factor 2 (SERF2)	10	9	2						
CGTTCCTGC	Id1*	10	5	0			T pool			
CCTGTAATG	exonuclease homolog RAD1 (S. pombe)	7	4	0						
ATGGTGGGG	zinc finger protein homologous to murine Zfp-36	4	4	0						
CACCTGTAA	homolog of Xenopus Claspin	8	0	0			T either			
GAAATGATG	c-myc binding protein	1	6	0						
ACCCCCCCG	junD	4	1	0						
	Growth factors and signal transduction									
TGGAGTGGA	guanylate kinase 1	7	15	1			T pool			
CCACTGCCC	Homo sapiens cDNA FLJ11978 fis, clone HEMBB1001271, similar to BUB1_human mitotic checkpoint serine/threonine-protein kinase	14	4	2			T either			
CTGGCCCTT	hepatocyte growth factor	10	1	6					T spec	N spec
AGCCCGGGA	oligophrenin-1 (GTPase regulator associated with focal adhesion kinase pp125(FAK))	7	5	2					T spec	
TCACTGCAC	FLJ14058, similar to I78885 serine/threonine-protein kinase	11	2	0			T pool			
TAGGTTGTC	tumor protein, translationally-controlled (histamine-releasing factor)	8	1	2						
CCATTGCAT	fibroblast growth factor receptor	6	4	1						
GGAGGTGGG	granulin (epithelin)	3	0	8				N pool		
TTGCCCCCG	AXL receptor tyrosine kinase	1	0	0						

					20 most abundant tags		Differential expression		Tissue-specific expression	
Tag	Gene match	T1	T2	N	T	N	T	N	T	N
	Protein folding, translocation and degradation									
AGAATCGCT	coatomer protein complex, subunit alpha	5	11	0			T pool			
CCCAAGCTA	heat shock protein 27	2	2	9						
TGCATCTGG	heat shock 70kD protein 5 (glucose-regulated protein, 78kD)	3	0	9				N pool		
GTGCTGGAC	proteasome activator hPA subunit beta	4	6	1						
GAGGTCCCT	proteasome (prosome, macropain) subunit, alpha type, 6	3	7	0			T either			
ACCCTTCCC	signal sequence receptor, beta (translocon-associated protein beta)	4	4	0						
GGCTGGGCC	clathrin, light polypeptide	0	1	7				N pool		
	Thymosins									
TTGGTGAAG	thymosin beta 4	17	20	17						
GGGGAAATC	thymosin beta 10*	10	34	4			T pool			
TCAGACGCA	prothymosin alpha*	13	7	2						
	Cytoskeletal proteins									
GGCTGGGGG	profilin	13	8	50		N high				
CGAGGGGCC	actinin, alpha 4	3	5	25				N pool		
CCTCCAGCT	keratin 8	2	0	13				N pool		
TCACCGGTC	gelsolin	5	1	8						
GAAGCAGGA	non-muscle cofilin	1	1	11				N pool		
GCCACGTGG	villin-like protein	0	2	9				N pool		
TCAGGGCTG	actin, gamma 1	10	0	0			T either			
ACAGCGGCA	desmoplakin	0	0	6				N pool		
	Other widely expressed general cellular genes									
CCCTGGGTT	ferritin, light polypeptide	11	10	11						
CCAAGGTGG	peptidylprolyl isomerase A (cyclophilin A)	20	0	0			T pool		T spec	
GCGACCGTC	aldolase A	6	9	9						
TACCATCAA	glyceraldehyde-3-phosphate dehydrogenase	1	2	0						
	Mucins, trefoil peptides and glycoprotein-related genes									
CCTGGGAAG	MUC1	11	3	18						
GTGATCAGC	MUC5*	2	9	43		N high		N pool	T spec	N spec
GCGAACGTG	MUC6	5	1	3						
CTGGCCCTC	pS2 (TFF1: trefoil factor 1)*	127	21	170	T1 high	N high			T spec	
AAATCCTGG	spasmolytic polypeptide (TFF2: trefoil factor 2)*	3	1	50		N high		N pool		N spec
CTCCACCCG	intestinal trefoil factor (TFF3: trefoil factor 3)*	13	1	0			T pool			
GGAAAACAG	galectin 4	0	5	11						
GGGCCTGGG	fucosyltransferase 6 (alpha (1,3) fucosyltransferase)	1	7	6						
	Immune defence									
CTCCCCCAA	immunoglobulin alpha, IgA (VPS28 or SNC73)	186	18	781	T1 high	N high				N spec
CTCCCCAAA	immunoglobulin alpha, IgA	1	0	18				N pool		N spec
CTCCCCAAG	similar to predominant tag for IgA or gamma-glutamylcysteine synthetase	1	0	28				N pool		N spec
CTCCCCCAC	similar to predominant tag for IgA or ESTs	0	0	8				N pool		N spec
GCGGAGGTG	immunoglobulin alpha, IgA (variable region)	31	6	40					T spec	N spec
TTCCCCCAA	immunoglobulin alpha, IgA (last NlaIII site but variant sequence)	9	0	38				N pool		N spec
AAGGGAGCA	immunoglobulin lambda light chain	31	18	215		N high		N pool		N spec
AAACCCCAA	immunoglobulin lambda gene cluster or ESTs	1	0	13				N pool		
AAAGGAGCA	immunoglobulin lambda gene cluster or ESTs	0	0	8				N pool		N spec
GAAACCCCA	immunoglobulin kappa light chain VJ region	5	9	1						
GTACGTATT	immunoglobulin J (joining) chain	104	1	55	T1 high	N high			T spec	N spec
GAAATAAAG	immunoglobulin gamma 3	4	3	52		N high		N pool		
CAAACTAAC	immunoglobulin mu	3	0	26				N pool		N spec
GTGCGCTGA	major histocompatibility complex, class I, C	0	0	18				N pool		
TGCAGCACG	major histocompatibility complex, class I	6	9	2						
AACGCGGCC	macrophage migration inhibitory factor	13	5	5						
ATGTAAAAA	lysozyme*	238	3	101	T1 high	N high			T spec	N spec
TGCCCTCAG	lipocalin 2 (neutrophil gelatinase associated lipocalin, NGAL)*	10	1	26						N spec
	Proteases and anti-proteases									
AGTGCTCTT	pepsinogen C	37	2	19					T spec	N spec
AACCTCCCC	pepsinogen A	0	14	0			T pool		T spec	
GCGCTGGAG	salivary alpha amylase	7	4	4						
AAGGTAACA	serine protease inhibitor, Kazal 1 (pancreatic secretory trypsin inhibitor)	5	3	6						N spec
TGCAGTCAC	collagenase I	10	2	0			T either			
GGAAAAGTG	alpha-1-antitrypsin	4	6	0						
GGCCAGGTG	procollagen (type III) N-endopeptidase	1	2	7						
ATGAGCTGA	cystatin B	1	2	0						
CCCCAGTTG	calpain 4, small subunit (30K)	1	0	7				N pool		

Tag	Gene match	T1	T2	N	20 most abundant tags		Differential expression		Tissue-specific expression	
					T	N	T	N	T	N
	Detoxification									
GGCCCAGGC	aldehyde dehydrogenase 3 family, member A1	11	1	20					T spec	N spec
TCTGTAATC	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1	4	7	0			T either			
CACTGGCAA	sulfotransferase family, cytosolic, 1C, member 1	1	0	8				N pool		N spec
CCAACCGTG	glyoxalase I	2	0	8				N pool		
AGGTCCTAG	glutathione S-transferase pi	6	4	6						
	Transport of water, ions and lipid									
CGTGGGGCT	aquaporin 5	8	1	8						N spec
CTTCCCCCA	proteolipid protein 2 (colonic epithelium-enriched) (PLP2)	2	0	12				N pool		N spec
TGGCCCCAG	apolipoprotein C-I	0	0	6				N pool		N spec
TGGTTGGTG	plasmolipin	0	0	6				N pool		
	Gastrin, new gene CA11 and PSCA									
GCTGGAGGA	gastrin*	0	0	133		N high		N pool		N spec
GCTGGAGGG	similar to predominant tag for gastrin or ESTs	0	0	11				N pool		N spec
TCATTCTGA	new gastric gene CA11 (= gene name, not carbonic anhydrase)*	0	1	91		N high		N pool		N spec
TCATTCTGG	similar to predominant tag for CA11 or no match	0	0	13				N pool		N spec
TCATTTTGA	similar to predominant tag for CA11 or no match	0	0	11				N pool		N spec
GCCCAGCAT	prostate stem cell antigen*	3	1	24				N pool		N spec
	Other genes expressed in normal stomach									
GCCGGGTGG	basigin (EMMPRIN)	2	15	11						
AAGGGAGCC	olfactory receptor, family 1, subfamily E, member 1	0	2	11				N pool		N spec
CAGGCCCCA	S100 calcium-binding protein A11 (calgizzarin)	3	0	12				N pool		
GCCTCCTCC	muscle specific gene M9 (despite name, tag expressed at similar levels in breast and pancreatic cancer cell lines)	1	0	7				N pool		
TGGCCATCT	PP1201 protein (N-methyl-D-aspartate receptor-associated protein)	1	2	9				N pool		
	Tissue-specific genes more highly expressed in tumors									
CCCCCTGCA	mesothelin	9	2	0			T either			
GGCCCTAGG	butyrate response factor 2 (EGF-response factor 2)	7	6	1						
TACCTCTGA	S100 calcium-binding protein P	36	2	13					T spec	
	Hypothetical proteins									
TCCCTATTT	hypothetical protein PRO2214	7	2	14						N spec
CACCCACTG	Homo sapiens HSPC323 mRNA, partial cds	0	1	8				N pool		N spec
CGCCTGTAG	Homo sapiens cDNA: FLJ21521 fis, clone COL05880	8	0	0			T either			
CTACTGCAC	hypothetical protein FLJ13087	10	4	0			T pool			
	Multiple matches									
GTGAAACCC	multiple matches	212	178	43	Ts high	N high	T either			
CCTGTAATC	multiple matches	210	141	15	Ts high		T pool			
CCACTGCAC	multiple matches	192	100	84	Ts high	N high				
GCGAAACCC	multiple matches	72	35	24	T1 high					
AGGTCAGGA	multiple matches	64	52	6	Ts high		T pool			
AACCCGGGA	multiple matches	67	29	8	T1 high		T pool			
CCATTGCAC	multiple matches	60	22	19	T1 high					
CCTGTAGTC	multiple matches	41	42	5	T2 high		T pool			
AACCCAGGA	multiple matches	31	23	6			T either			
GCAAAACCC	multiple matches	40	16	4			T pool			
GTGGCTCAC	multiple matches including KIAA0414 protein	14	32	7			T either			
GTGGTGGGC	multiple matches including DKFZp761F152	19	24	5			T either			
CCACTGTAC	multiple matches including FLJ12320 fis, clone MAMMA1002082	23	7	4			T either			
GAGAAACCC	multiple matches	17	15	2			T pool			
AGCCACCGC	multiple matches including RAB, RAS oncogene family-like 2B	17	11	2			T pool			
GTGAAACTC	multiple matches including KIAA0328 protein	11	12	2			T either			
TTGGTCAGG	multiple matches	6	17	1			T pool			
CCTATAATC	multiple matches	7	16	0			T pool			
CCTGTGGTC	multiple matches	15	7	1			T pool			
GTGGTGCGC	multiple matches	11	8	1			T pool		T spec	
AGCCCAGGA	multiple matches including FLJ10940 fis, clone OVARC1001162	12	4	2					T spec	
CCACTACAC	multiple matches	11	6	0			T pool			
GTAAAACCC	multiple matches	6	10	0			T pool			
AGCCACTGT	multiple matches	8	7	0			T pool			
AGGCTGAGG	multiple matches including FLJ14137 fis, clone MAMMA1002764	5	6	1					T spec	
GAAACTGAA	multiple matches including eukaryotic initiation factor 2-associated protein, p67	1	10	0			T either			
GGCGACAGA	multiple matches including RAB3B, member RAS oncogene family	8	1	0			T either			
CAAAAAAAA	multiple matches including ovarian carcinoma immunoreactive antigen	0	1	7				N pool		
TTTTTTTTT	multiple matches including Homo sapiens clone 23620 mRNA sequence	1	0	6						N spec
GAAATAAAA	multiple matches including FLJ21286 fis, clone COL01915	0	0	6				N pool		

Tag	Gene match	T1	T2	N	20 most abundant tags		Differential expression		Tissue-specific expression	
					T	N	T	N	T	N
	No matches									
TCCCTATTA	non-clustered ESTs only, tag also high in prostate	67	30	120	T1 high	N high				
TCCCCGTAC	ESTs, tag moderately high in prostate and colon	68	20	109	T1 high	N high				
ACGCAGGGA	matches ESTs only for histone deacetylase 3 and glucose phosphate isomerase (NOT heat shock protein 90*), tag moderately high in prostate and colon	89	79	6	Ts high		T pool			
TCTCCATAC	Homo sapiens, clone MGC:10923, mRNA, complete cds (no Unigene cluster), tag also high in prostate	76	0	0	T1 high		T pool		T spec	
TTGGTTGGC	weakly matches a few ESTs only in Unigene cluster for glycoprotein (transmembrane) nmb, tag essentially absent from other SAGE libraries	0	0	34				N pool		N spec
GTGACCACG	weakly matches ESTs only in Unigene cluster for glutamate receptor, N-methyl D-aspartate 2C, tag moderately high in breast and colon	22	10	0			T pool			
TGTGTTGAA	no match, tag low in other SAGE libraries	11	13	6					T spec	N spec
GAAGCCCCA	no match, tag at similar levels in malignant mesothelioma SAGE library	22	3	6						N spec
TCCCCGTAT	no match, tag low in other SAGE libraries	11	0	18						N spec
AATTAAATT	weakly matches ESTs	28	0	0			T pool		T spec	
GGGAAGCAG	matches EST only in Unigene cluster for FLJ12175, clone MAMMA 1000713, weakly similar to L-ribulokinase; tag high in breast	2	0	26				N pool		
CTTCCCCAA	no match, tag low in other SAGE libraries	8	0	18						N spec
CTCCCCCAG	no match, tag low in other SAGE libraries	7	0	19				N pool		N spec
TTCTTGTGG	matches many ESTs only in Unigene cluster for ribosomal protein S11	1	18	1			T pool			
AACCCGGGG	no match, tag lower in other SAGE libraries	10	5	4					T spec	
TCCCCGTAG	no match	2	0	17				N pool		N spec
ATGGCAAGG	matches ESTs only for DR1-associated protein 1 (negative cofactor 2 alpha)	5	10	1			T either			
CCACTTGCA	no match, tag low in other SAGE libraries	8	2	6						N spec
CCGACGGGC	no match	1	13	2			T either			
ACGCAAGGA	weakly matches ESTs	6	8	1					T spec	
CTCCCCCCA	no match, tag essentially absent from other SAGE libraries	3	0	12				N pool		N spec
GTACCGTAT	no match, tag essentially absent from other SAGE libraries	9	0	6						N spec
TTTAGGATG	ESTs in Unigene clusters Hs.128713 and Hs.156077 (both contain only ESTs), tag essentially absent from other SAGE libraries	0	0	15				N pool		N spec
TCCGCGAGA	matches ESTs only for zinc finger protein homologous to Zfp-36 in mouse (second tag?)	3	11	0			T pool			
CCATCGTCC	no match	3	10	1			T either			
AAGGGAGTT	no match, tag lower in other SAGE libraries	4	0	8						N spec
CTGGCCCTG	no match, tag low in other SAGE libraries	4	0	8						N spec
AAGGGAGCG	no match, tag essentially absent from other SAGE libraries	3	0	8				N pool		N spec
GACTCTGGT	no match, tag at similar levels in breast and prostate carcinoma	3	0	8				N pool		
GCCATCCCC	ESTs and possible match to mitochondrial sequence	1	10	0			T either			
TCCCGGTAC	no match, tag at similar levels in mesothelioma and ovarian carcinoma	3	0	8				N pool		N spec
CTCGTTAAG	no match (Unigene)	3	7	0			T either			
GAGGGAGCA	no match, tag low in other SAGE libraries	2	0	7						N spec
ATGATGGCA	no match	0	8	0			T either			
GCTGGAGGT	no match, tag essentially absent from other SAGE libraries	0	1	7				N pool		N spec
CCCCCCCAA	no match, tag essentially absent from other SAGE libraries	1	0	6						N spec
GGGTTGGGG	no good match, tag at similar levels in prostate carcinoma SAGE libraries	0	0	7				N pool		N spec
GCTGGAGAA	no match, tag low in other SAGE libraries	0	0	6				N pool		N spec
GTGATCAGT	no match, tag essentially absent from other SAGE libraries	0	0	6				N pool		N spec
TCATTCTGC	no match, tag essentially absent from other SAGE libraries	0	0	6				N pool		N spec

Table 4.3: Immunohistochemistry for selected genes in a range of gastro-intestinal tissues

This table summarises the results of immunohistochemical staining for candidate genes in a range of gastro-intestinal tissues. Further information about the gastric tumours is presented in Table 2.1. All four cell lines are derived from adenocarcinomas (see Section 2.1.5.3).

The + and – symbols are used to indicate the intensity of the staining, where – means no staining, +/– means focal weak staining, and +++ means the strongest positive staining. Figure 4.2 shows immunohistochemical staining in the tissues used for SAGE. Spasm polypep indicates human spasmolytic polypeptide (TFF2).

Tissue or cell sample		MUC5	pS2	Spasm polypep	Lysozyme
gastric tumour 1	distal, intestinal	+	+++	–	+++
gastric tumour 2	distal, intestinal	++	+/–	–	++
gastric tumour 3	distal, intestinal	+/–	+	–	+
gastric tumour 4	distal, diffuse	++	+	–	++
gastric tumour 5	distal, diffuse	++	+	+++	+++
gastric tumour 6	proximal, intestinal and solid	+	+	+	+
gastric tumour 7	proximal, diffuse and intestinal	–	+	–	–
gastric tumour 8	proximal, mixed	–	–	–	+/–
AGS	gastric adherent cells	+	++	–	+++
KATO-III	gastric spherical cells	–	+/–	–	++
MKN-45	gastric spherical cells	–	+	–	++
OE-19	esophageal adherent cells	–	+	–	++
stomach	normal antrum	+++	+++	+++	+++
stomach	normal body	+++	+++	+++	++
stomach	normal cardia	+++	++	+++	+++
oesophagus	normal squamous mucosa	–	–	–	+
oesophagus	squamous carcinoma	–	–	–	–
small intestine	normal	–	–	–	++
colon	normal mucosa	–	+/–	–	–
colon	adenocarcinoma	+	+++	–	++

5 Further investigation of new gastric gene

5.1 Results

As described in the previous chapter, one tag, TCATTCTGA, was very interesting. It was expressed at very high levels in normal stomach, but was absent from the gastric carcinomas, and comparison with the other glandular epithelial libraries showed that this tag was gastric-specific. When it was first noticed, this tag did not match to any known genes in the genetic databases, although subsequently the gene has been described in a single publication (Yoshikawa *et al.* 2000). In fact, this tag was one of the ten most abundant transcripts in the normal stomach: the others included gastrin, immunoglobulin alpha, lysozyme, gastric mucin (MUC5) and trefoil peptides. These other highly expressed genes are, not surprisingly, important in normal gastric function, suggesting a similar role for the unmatched transcript.

In this chapter, the matching transcript is first identified. The expression profile of the gene is then evaluated by Northern blotting and *in situ* hybridisation. The cDNA itself is further characterised, not only in its human form, but also in other species. A genomic clone is identified and used to map the gene to human chromosome 2. The genomic sequence for both the human and mouse is identified and compared. The gene at its mRNA and DNA levels is highly homologous between species. The predicted protein sequences are presented and compared. Lastly I speculate on the possible function of this gene and methods by which this could be further investigated.

5.1.1 Investigation of the new gastric gene at the RNA level

5.1.1.1 EST analysis

At the time of first identification, investigation through SAGEmap showed that this tag did not correspond to any known gene in Genbank. However, it did match ESTs, within a single human (Homo sapiens) gene cluster on NCBI's UniGene database (Schuler 1997). At the time the cluster contained only eight EST sequences; this has since increased and Figure 3.1 shows the web-page of the corresponding current UniGene cluster. The eight ESTs were down-loaded from the web then analysed using the DNASTAR suite of programs, into which the sequences were imported through EditSeq.

Using the sequence assembly program SeqMan, it became obvious that the eight ESTs in fact overlapped, and they were merged to create a single likely cDNA sequence with a length of around 750 bp, as shown in Figure 5.2. The cDNA sequence was found to contain a start codon followed by a long open reading frame (ORF, or translatable protein sequence), suggesting that this represents a true protein-coding cDNA.

The cDNA sequence has been further characterised and details of this are presented later in Section 5.1.1.10 and thereafter. Initially, however, the sites of expression of the mRNA were studied. As expected, six of the original eight on-line ESTs originated from gastric tissue. Two of the ESTs, however, originated from a library comprising a pool of tissues from heart, melanocytes and pregnant uterus. It would of course be possible that this gene is also expressed in the heart, in melanocytic cells, or in the normal uterus. Although none of these three bear an obvious similarity to the stomach in terms of likelihood of shared tissue-specific gene expression, the uterus, like the stomach, is a glandular and thus secretory organ. An alternative explanation could be that the tissue labelled "pregnant uterus" from which the cDNA libraries of interest were created also contained foetal tissue, possibly including stomach, or placenta. I tried to find out whether the libraries contained only uterine tissue (the lining, the endometrium; and/or the uterine muscle coat, the myometrium) or also placental or foetal tissue, but the scientists who supplied the libraries to NCBI did not know these details, saying simply that the tissues had been so described by their pathology colleagues (personal communication, Marcello Bento Soares).

5.1.1.2 Creation of a cDNA probe for the new gastric gene by RT-PCR

Complementary single-stranded oligonucleotide primers were designed using the Vector NTI suite of programs. Reverse transcriptase polymerase chain reaction (RT-PCR) was then performed on a normal gastric mRNA sample. First, a small central cDNA fragment of around 100 bp was amplified, to confirm that the predicted central overlapping sequence was correct. Then, a near full-length cDNA fragment was amplified, purified and cloned into Invitrogen's TOPO-TA vector. In addition, the IMAGE clones of the ESTs of the new gastric gene were obtained from HGMP in Cambridge (Lennon *et al.* 1996). Appropriate larger quantities of DNA were obtained from the clones by plasmid preparation. All clones were sequenced using appropriate primers (M13 forward and reverse or Sp6, T3 or T7) for their flanking polylinkers to confirm their identity prior to use. These clones were then used to prepare probes for initial investigation of the expression profile of the gene through Northern blotting and *in situ* hybridisation.

5.1.1.3 Northern blotting for the new gastric gene

Northern blotting is a technique which identifies mRNA species that have been size-fractionated by gel electrophoresis. A probe for the new gastric gene was created by radio-labelling the near full-length cDNA fragment. Using the same Northern blots described in the previous chapter, this probe gave rise to a clean, single band running at approximately 1 kb. This estimated size fits well with the predicted cDNA length of 750 bp plus a poly(A) tail, which on average contains around 200-250 adenylate residues (Strachan *et al.* 1999). As shown in the previous chapter, the new gene was expressed only, but at high levels, in normal stomach, in all of its anatomical areas, from the cardia though the body to the antrum (Figure 5.3). mRNA for the new gastric gene was absent from all of the other gastro-intestinal samples, including: gastric adenocarcinomas and gastro-oesophageal adenocarcinoma cell lines; normal squamous oesophagus, small intestine and colon, and their corresponding tumours, oesophageal squamous carcinoma and colonic adenocarcinoma (Figure 5.3).

The probe was then applied to two more, commercial, Northern blots: Invitrogen's Northern Territory™ Human Normal Tissue Blot III, and OriGene Technologies' Multiple Choice™ Northern Blot, as shown in Figure 5.4. Once more, foveolin was expressed only in normal stomach. Foveolin mRNA was absent from other gastro-intestinal tissues, including colon, appendix, gallbladder and liver. It was also absent from lymphoid organs, including tonsil, thymus, lymph node and spleen, and from genito-urinary organs including prostate, testis, ovary and placenta. Unfortunately, given the previous information from the on-line libraries, no uterine sample was present on these blots for assessment of possible expression of the new gastric gene.

5.1.1.4 Comparison of expression profiles with the Japanese study

After I had obtained these initial results, a single publication on this new gene emerged (Yoshikawa *et al.* 2000). This Japanese research group had also identified the gene as being expressed more highly in normal stomach than in gastric carcinoma, through an alternative expression profiling technology called differential display. They then went on to profile the gene's expression pattern in various tissues. Using a PCR-based method called Rapid Amplification of cDNA Ends (RACE) applied to a commercial panel of cDNAs, mRNA for the new gene was identified at high levels in stomach, as expected but also, at lower levels, in placenta and uterus. (I also performed RACE, but with the aim of characterising the 5' and 3' ends of the cDNA rather than for expression profiling purposes. This is discussed later in Section 5.1.1.10.) In contrast, by Northern blotting this group identified the mRNA in stomach only, not in placenta; again, their commercial blots lacked a uterine sample.

Thus my data, the on-line information, and the Japanese study agreed that the new gene was expressed very highly in stomach. Although its expression outwith the stomach seemed to be limited, the on-line information and Japanese study seemed to suggest that it might also be expressed in uterus and, possibly, placenta. I therefore went on to examine the expression and localisation of the mRNA for the new gastric gene in tissues using a second method, *in situ* hybridisation.

5.1.1.5 *In situ* hybridisation (ISH) for the new gastric gene in normal stomach

In situ hybridisation (ISH) is a technique which identifies and localises mRNA within individual cells or tissue sections (McNicol *et al.* 1997). For ISH, a digoxigenin-labelled riboprobe (ribonucleotide probe) was prepared by *in vitro* transcription of the cloned near full-length cDNA fragment, using RNA polymerase sites on the polylinker at either end of the insert sequence.

In the positive control slides, the mRNA was clearly and cleanly localised by ISH to the superficial/foveolar zone of normal gastric mucosa, as shown in Figure 5.5 (see also Section 1.2.2). In normal gastric body mucosa, the superficial/foveolar epithelium occupies the upper quarter of the gastric glands; the superficial zone merely means the surface while the foveolae are simply the immediately adjacent gastric gland openings, or, literally, pits. The deeper parts of the gastric glands comprise the proliferative zone, which occupies the next quarter of the glands, and the gastric pits, which make up the lower half of the glands, and the mRNA was absent from these areas. Staining in gastric antral and cardiac mucosa was similarly limited to the superficial/foveolar zone which occupies the upper third of the glands in the antrum and around the upper half of the glands in the cardia (data not shown). Staining was absent from the negative control slides (Figure 5.5).

As discussed later in Section 5.1.1.10, BLAST sequence searches indicate that homologues of foveolin exist in other mammalian species, specifically in mouse, rat and cow. This riboprobe was therefore also applied to normal stomach from these species. (For the cow, the fourth stomach, the abomasum, was used: this is the most similar to the single stomach in non-ruminant mammals such as humans). Even though the probe used was for the human mRNA, positive staining was present in the superficial/foveolar epithelium in these other species, in a location identical to that in humans. Staining in the mouse gastric mucosa is shown in Figure 5.5.

5.1.1.6 What's in a name: CA11 or foveolin?

The Japanese group had called the gene CA11 (Yoshikawa *et al.* 2000). The GenBank Accession Numbers for the two mRNAs are: NM_019617 and AB039886. However, this seemed an odd choice of name, since CA11 is already the accepted gene symbol for carbonic anhydrase type XI, which bears no relationship or homology to the new gene, by either UniGene clustering or BLAST analysis. Presumably the name related to the number of bands in their differential display or to the sample number. However, given its site of expression in the gastric mucosa, I have instead provisionally termed this gene foveolin.

5.1.1.7 *In situ* hybridisation (ISH) for foveolin combined with immunohistochemistry in normal stomach

Immunohistochemistry (IHC) is a technique which identifies and localises protein within individual cells or tissue sections, using specific antibodies. ISH for foveolin was combined with IHC, as described in the previous chapter, to compare its localisation with that of the other highly expressed normal gastric genes.

Gastric mucin (MUC5) and pS2 (trefoil factor 1 (TFF1)) were present in the superficial/foveolar epithelium in a similar though not identical location to that of foveolin (Figure 5.5). MUC5 and pS2 are usually co-localised and, outwith the stomach, are widely expressed in the gastro-intestinal tract and in other glandular epithelial organs including lung, breast, prostate, ovary and uterus (Williams *et al.* 1997). The similar locations of foveolin, MUC5 and pS2 in the stomach would provide support for the possibility that foveolin could also be present in these other tissues.

5.1.1.8 *In situ* hybridisation (ISH) for foveolin in other normal and diseased upper gastro-intestinal tissues

As already described, foveolin is highly expressed in the superficial/foveolar epithelium in all areas of normal stomach, but is absent from gastric carcinoma. The study of its expression was then extended to a wide panel of other normal and diseased gastro-intestinal tissues including gastritis, duodenal ulcer and Barrett's oesophagus (see Section 1.3.)

In the stomach, foveolin expression was essentially unchanged in areas of gastritis, including that caused by *H. pylori* infection, or of gastric atrophy alone (see Section 1.3.1). Not surprisingly, in areas of ulceration, where by definition the epithelium is denuded, foveolin was absent. However, foveolin was also absent from areas of intestinal metaplasia (Figure 5.6) and dysplasia (Figure 5.6), as well as from carcinoma. As clearly seen in the figures, the change in phenotype from normal gastric epithelium to metaplastic intestinal or dysplastic epithelium was extremely abrupt, with an accompanying abrupt change in the presence or absence of staining for foveolin mRNA.

In the oesophagus, foveolin mRNA was found only in areas of gastric metaplasia, occurring within the generalised metaplastic glandular mucosa present in Barrett's oesophagus (see Section 1.3.2). It was not present in normal oesophageal squamous mucosa or normal submucosal glands, nor was it found in intestinal metaplasia in Barrett's oesophagus or in oesophageal squamous carcinoma or adenocarcinoma.

In the duodenum, foveolin mRNA was found only in small areas of gastric metaplasia (Figure 5.6), occurring at the tips of the small intestinal villi. Gastric metaplasia in the duodenum is a well-recognised phenomenon, which occurs in response to epithelial damage and ulceration in the presence of excess acid, usually in association with an antral-predominant *H. pylori*-associated chronic gastritis (see Section 1.3.1). Foveolin was not present in normal duodenal mucosa or submucosal Brunner's glands, nor was it found in duodenal ulcers, except in focal gastric metaplastic epithelium at the ulcer edge.

Elsewhere in the small and large intestines, foveolin mRNA, not surprisingly, was present in ectopic gastric mucosa present in Meckel's diverticula, which are congenital anomalies present in around 2% of the population. The Meckel's diverticulum is a small pouch in the small intestine; around half contain ectopic body-type, hence acid-secreting, gastric mucosa (Cotran *et al.* 1994). Foveolin mRNA was also present in gastric pyloric metaplasia in the colon, which is a relatively rare phenomenon seen in inflammatory states such as inflammatory bowel disease (Cotran *et al.* 1994): this regenerative gastro-intestinal epithelium is now known as the ulcer-associated cell lineage (UACL) (Wong *et al.* 1999). Foveolin was therefore only expressed in the gut in native or metaplastic gastric epithelium.

5.1.1.9 *In situ* hybridisation (ISH) for foveolin in tissues outwith the gastro-intestinal tract

Having established that within the gut, foveolin is essentially specific to epithelium of gastric type, the next aim was to investigate whether foveolin is expressed outwith the gut, particularly in the uterus, placenta and other gynaecological tissues, as suggested by the Japanese study and on-line cDNA libraries, or in other glandular epithelia, as could be postulated on the basis of its similar location to MUC5 and TFF1.

ISH for foveolin mRNA was performed in a wide range of normal and abnormal gynaecological, glandular epithelial and other tissues, as listed in Section 2.1.5.2. Foveolin mRNA was found only in a subset of ovarian mucinous tumours, both benign (cystadenomas) and malignant (adenocarcinomas) (data not shown). No staining was found in any of the other tissues, and in particular, foveolin mRNA was not identified in any of the five uterine and three placental samples examined. The uterine tissue sections included normal endometrium, in proliferative and secretory phases of the menstrual cycle, and myometrium, as well as endometrial adenocarcinoma.

5.1.1.10 Characterisation of human foveolin mRNA by 5'- and 3'-rapid amplification of cDNA ends (RACE)

Having investigated where foveolin mRNA was expressed, I now wished to characterise the mRNA itself in more detail. Rapid amplification of cDNA ends (RACE) is a common method used to analyse, separately, the 5'- and 3'-ends of mRNAs. 5'- and 3'-RACE for human foveolin was therefore performed using Clontech's kit. Oligonucleotide primers were designed using Vector NTI in accordance with advice given in the kit. For both 5'- and 3'-RACE reactions, PCR produced a single clear product band (data not shown). The PCR products were then purified and sequenced five times. The sequence files were analysed and assembled using SeqMan.

A composite diagram of these RACE results is presented in Figure 5.7. My 3' RACE results for human were identical to the results of the Japanese study and to the on-line, and locally sequenced, ESTs, which now numbered more than eight.

The 5' RACE results were broadly similar to the on-line EST sequences. However, even the 5' most of the on-line ESTs started at position 75 (CTCCTC...). In contrast, my 5' RACE included a further 11 bp upstream (ATGCTT...). In fact, all of my 5'-RACE sequences contained an additional 5' ACGCGGGG, which was absent from both the sequence presented in the Japanese study and from the genomic sequence which is described later in Section 5.1.2.4; this is likely to represent artefact from the SMART™ (Switching Mechanism At 5' end of the RNA Transcript) cDNA synthesis methodology used in the Clontech RACE kit. In the Japanese study, although most 5' RACE species were of similar length to mine (their primers were different and resulted in fragments of 79 bp), the longest sequence, which was preferred, and was used for publication, included an additional upstream 50 bp (this fragment was thus 132 bp long) (Yoshikawa *et al.* 2000). As the Japanese group explained in their paper, the 5' end of CA11 has some self-complementary sequences: because of the strong secondary structure, most cDNAs were thought to have lost their 5' ends during first-strand cDNA synthesis (Yoshikawa *et al.* 2000). This extra 5' sequence contains an additional, in-frame, upstream ATG, that is, translation start site, which obviously would provide an alternative protein coding sequence.

In order to try to ascertain which ATG was likely to be functional, I went on to examine the foveolin mRNAs of other species and also to investigate its genomic sequence.

5.1.1.11 Characterisation of mouse, rat and bovine foveolin mRNA by 5'- and 3'-rapid amplification of cDNA ends (RACE) and analysis of on-line sequences

The sequence of human foveolin mRNA was used with NCBI's BLAST® (Basic Local Alignment Search Tool) to search for homologous sequences from other organisms. The programs used were: BLASTN, which compares a nucleotide query sequence against a nucleotide sequence database; and TBLASTX, which compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database (Wheeler *et al.* 2001). The default parameters were used. The nucleotide sequence databases used were: NR, which contains all characterised sequences from GenBank and other sources (EMBL+DDBJ+PDB); and dbEST which contains sequence data and other information on ESTs, from a number of organisms. Matching sequences were found in the databases from mouse and rat and, more recently, from cow.

Although initially the number of matching mouse sequences was small, they now number in their hundreds and constitute a single UniGene cluster, Mm.46414 (Mm = Mus musculus), for what has been named the RIKEN cDNA 2200002K21 gene. Most of the 10 mRNA/Gene sequences and 478 ESTs originate from the RIKEN Mouse Gene Encyclopaedia Project, which is a Japanese effort to collect and sequence full-length complementary DNAs expressed in mouse tissues (Kawai *et al.* 2001). The mouse strain used throughout is C57BL/6J. The reference gene is Accession Number NM_025466.

This mRNA was highly represented in two mouse gastric libraries, constituting 5.7% of library 489 (RIKEN full-length enriched, adult male stomach) and 4.0% of library 217 (Mus musculus stomach C57BL/6J adult). Most of the on-line mRNAs and ESTs originate from stomach. The tongue is the origin of 1 out of the 10 mRNAs and 27 of the 478 ESTs. Other EST sources include: 16 from pancreas, 17 from small intestine, 2 from caecum (the proximal colon) and 1 from ovary/uterus. The chromosomal location remains uncharacterised.

When the initial small number of matching mouse ESTs were identified, they were down-loaded from the NCBI web-site, assembled using SeqMan and used to create a putative single cDNA sequence. This was then used to design primers for 5'- and 3'-RACE, as before. mRNA was extracted from a sample of mouse stomach and RACE was performed, as before. This sequence is presented in Figure 5.7. The mouse RACE sequence was compared with the assembly of the down-loaded, more recently available sequences. The RACE reaction proved to be efficient and provided an additional 8 bp at the 5' end compared to the longest of the on-line sequences. However, at the 3' end, 7 out of 380 ESTs or mRNA sequences which assembled together, including the reference sequence NM_025466 which is 1220 bp long, ran significantly beyond the 3' end of the

RACE sequence. This rare but long additional 3' sequence agreed with the mouse genomic sequence, which is described later in Section 5.1.2.5, and may represent a lack of cleavage beyond the 3' AAUAAA polyadenylation site.

In contrast to the numerous matching mouse sequences, only two rat sequences emerged, within the single UniGene Cluster Rn. 43048 (where Rn indicates Rattus norvegicus). The ESTs were AI639014 and AI639511 (which are the GenBank Accession Numbers), from a pooled mixed-tissue library from the Sprague-Dawley rat strain. One of the ESTs is a 5' read while the other is a 3' read, so that, although it appears similar to the mouse cDNA, most of the rat sequence is single-copy and unconfirmed, and I have not yet performed rat RACE. The rat sequence is therefore not included in Figure 5.7.

Matching cow sequences have appeared in the databases only recently. These ESTs are BG937912, BG937850, BG937720, BG938026, BG937628 and BG937739, all from a cDNA library from bovine abomasum. The latter is the fourth stomach in the cow and most closely resembles the human stomach. No UniGene cluster yet exists for these sequences, which were down-loaded and assembled as described for the mouse ESTs. Although cow RACE has not yet been performed, the number of bovine sequences provides significant overlap and hence reasonable confidence in their accuracy. The putative cow cDNA is therefore included in Figure 5.7.

No other convincing homologies emerged with mRNAs from other, including lower, organisms, by either BLAST searching of the sequence databases, or use of NCBI's HomoloGene database (Wheeler *et al.* 2001). The latter, however, lacked any curated records of the obvious links between this gene in the human, mouse, rat and cow databases; indeed, the only calculated orthologue was to a Drosophila protein called dec-1 (Drosophila melanogaster defective chorion-1 fc106 protein precursor mRNA) which showed no significant similarity using BLAST or other methods of comparison. HomoloGene may thus not yet be entirely reliable.

One reason for investigating these different species related to the debate about the 5' end of the mRNA, which was around 50 bp longer in the Japanese study than by my results. The longer mRNA yields an additional upstream (5'), and inframe, ATG, or start codon. It also agrees with the human genomic sequence, which is described later in Section 5.1.2.4. However, although the human, mouse and cow mRNA sequences are highly homologous around and downstream of the second ATG (position 106 in Figure 5.7), upstream towards the first ATG the homology is much less. The results of these mRNA comparisons are echoed in the three translated protein sequences, which are described later in Section 5.1.3. While the varying levels of homology between the mRNAs can be easily seen by eye-balling the sequences, they are confirmed by BLAST analysis. The issue of which ATG in the human foveolin mRNA is likely to be functional is further debated in the discussion (Section 5.2).

Having thus examined both the expression and structure of foveolin mRNA in humans and other species, the next aim was to investigate its genomic structure.

5.1.2 Investigation of foveolin at the DNA level

5.1.2.1 Identification of a human genomic clone through probing a genomic library

The aim was to identify a clone containing the human foveolin genomic sequence. The human PAC library RPCI1 is a genomic library within PACs (plasmid artificial chromosomes) (Osoegawa *et al.* 2001). The library is spotted on seven high-density gridded filters and was obtained from HGMP (the Human Genome Mapping Project at Cambridge). The previously described radiolabelled cDNA probe was used to probe the PAC libraries. Unfortunately, some of the blots had been hybridised before and blots 1-4 were found to have been irrevocably damaged by stripping. Luckily, the remaining intact blots (5-7) contained a single positive clone in blot 7, as shown in Figure 5.8. The clone identification number was 291-N4. This clone was obtained from HGMP and a DNA plasmid preparation was performed. The cloned PAC DNA was then used as a probe for fluorescent *in situ* hybridisation (FISH) and was also used for genomic sequencing.

5.1.2.2 Fluorescent *in situ* hybridisation (FISH) for the human foveolin gene

The UniGene cluster for the human gene describes a matching sequence tagged site (STS) (UniSTS entry: stSG31094) on Chromosome 2 (D2S292-D2S145). The genomic clone was therefore fluorescently labelled and used in fluorescent *in situ* hybridisation (FISH) along with chromosome paint for chromosome 2. The results are shown in Figure 5.9, which confirms the presence of two copies of the gene on chromosome 2, on the small arm close to the centromere, in agreement with the STS prediction. Significant genetic instability of this region has not been reported in the Mitelman Database of Recurrent Chromosome Aberrations in Cancer (Mitelman *et al.* 1997).

5.1.2.3 Sequencing of the partial human genomic clone

The genomic clone was then used to obtain genomic sequence. Initially, the primers used for the PAC sequencing PCRs were both forward and reverse primers from various sites along the length of the foveolin cDNA sequence. The resulting genomic sequences were assembled using SeqMan and intronic sequence was gradually built up and used to design further primers for walking backwards and forwards from the ends of the known sequence.

Using this method, around 4 kb of sequence from the 3' end and mid-portion of the gene was obtained through laboratory effort. After a while, however, it became obvious that although long lengths of further sequence had been obtained, the next exon had not yet been reached. Further investigation confirmed that much of this sequence was actually of PAC vector origin. Thus, although the insert size is on average 110 kb (Osoegawa *et al.* 2001), I had been unfortunate in obtaining a clone in which the sequence of interest was at the extreme end of the insert, and in identifying only one positive clone.

I did not pursue this further, since by then on-line databases containing human genomic sequences had become available, and these were used instead.

5.1.2.4 Use of the on-line Celera genomic databases to obtain the full-length human genomic sequence

During my work on the PAC library and its positive clone, I had regularly interrogated the public NCBI databases in hopes that the human genomic sequence should become available for foveolin, saving further sequencing. The databases in question included not only the annotated NR database of known genes but also the more preliminary HTGS and GSS databases, containing Unfinished High Throughput Genomic Sequences and Genome Survey Sequences, including single-pass genomic data, respectively (Wheeler *et al.* 2001). No match emerged from these NCBI databases. Since this public effort is said to cover about 94% of the human genome, this was distinctly unlucky (Lander *et al.* 2001).

The more recently available commercial Celera human genomic databases (Venter *et al.* 2001), however, have been more successful in this regard and yielded a large 50 kb segment of genomic sequence. When this was down-loaded and analysed, its 3' end was found to contain what appeared to be the full-length human genomic sequence of foveolin. This was then compared with the human mRNA sequence. The latter, together with the GT-AG rule, which means that introns almost always start with GT and end with AG (Strachan *et al.* 1999), enabled the identification of the exons and delineation of the intron-exon boundaries. Analysis of the Human Genome Sequence suggests that the GT-AG rule is followed in over 98% of confirmed introns (Lander *et al.* 2001). This annotated human genomic foveolin sequence is presented in Figure 5.10.

5.1.2.5 Use of NCBI's Trace Archive databases to obtain the full-length mouse genomic sequence

By this time, NCBI's on-line mouse genomic resources had become much more extensive, with Whole Genome Shotgun sequencing reads generated by the Mouse Sequencing Consortium available on-line via the Trace Archive at Ensembl. This database was interrogated with the mouse foveolin cDNA sequence. Numerous matching mouse genomic sequences emerged and were down-loaded and assembled using SeqMan. Sequence from the intronic ends of these assemblies was used to re-interrogate the databases. Eventually the full-length mouse gene was obtained, with the final assembly containing 68 separate sequences, of which a "strategy view" diagram is shown in Figure 5.11.

Once again, this full-length mouse genomic sequence was compared with the mouse mRNA, and using the GT-AG rule, the exons and exon-intron boundaries were identified. This annotated mouse genomic sequence is presented in Figure 5.12.

5.1.2.6 Comparison of the human and mouse genomic sequences

The human and mouse genomic sequences and their exon-intron structure were then compared and the results are presented in Figure 5.13. As can be seen, the structure of these two genes is very similar. Both contain six exons, of similar length in both species.

5.1.3 Investigation of the new gastric gene at the protein level: the predicted foveolin proteins

The Open Reading Frames (ORFs), or translatable protein sequences, within the three human, mouse and cow mRNAs were translated using EditSeq. The resulting protein sequences are presented and compared in Figure 5.14. The alignments were performed using NCBI's BLAST2 and Clustal W multiple sequence alignment at the EMBL European Bioinformatics Institute.

Because the cow mRNA sequence originates from ESTs, no other predicted protein sequence exists. This mouse protein sequence agrees with that predicted in the GenBank entry for the mouse reference mRNA sequence NM_025466. However, the human CA11 GenBank entry predicts additional upstream amino acids corresponding to the extra 50 bp and upstream ATG. As with the mRNA sequences (see Sections 5.1.1.10 and 5.1.1.11), the strong homology shown by the proteins is lacking in the upstream segment if the first human ATG is used. (Since there is less upstream information for the cow sequence, the comparison simply includes an in-frame translation of the cow sequence which exists.) If the upstream segment in the human CA11 reference sequence is omitted, however, the three protein sequences are highly homologous.

The primary structure of the proteins was then studied using various of the Network Protein Sequence Analysis programs on-line at the Pôle Bio-Informatique Lyonnais. Table 5.1 shows the physico-chemical properties of the foveolin proteins from human, mouse and cow. The human protein is that generated using the second ATG site. It is clear that in this respect also the three proteins are very similar. All three species are predicted to contain a starting signal peptide; but this applies to the human sequence only if the second ATG is used rather than the first. Partly based on the presence of such a signal peptide, the proteins are also predicted to be extracellular, being either expressed on the outer surface of the cell membrane or secreted.

The secondary structure of the three foveolin proteins was then predicted as a consensus of the results from twelve methods within the Network Protein Sequence Analysis programs on-line at the Pôle Bio-Informatique Lyonnais, as shown in Figure 5.15. Again, the three proteins are similar but unfortunately it is not possible to extrapolate from these secondary structures to arrive at the likely functions of the protein.

The predicted foveolin proteins were then compared against various protein databases using BLASTP which compares an amino acid query sequence against a protein sequence database and were further analysed with the Network Protein Sequence Analysis programs on-line. No similarities with any known proteins were detected, nor were there any obvious well-characterised protein motifs.

5.2 Discussion

Large-scale mRNA expression profiling of gastric carcinoma and normal stomach by SAGE yielded one particularly interesting unknown transcript which has been further investigated here. The characterisation of a gene involves its study at the DNA, mRNA and protein levels. At the RNA level, Northern blotting confirms that the new gene is highly expressed in normal stomach, in all three zones, namely the cardia, body and antrum, but is absent from gastric carcinoma. *In situ* hybridisation confirms the presence of this mRNA in the stomach, where it is located in the superficial/foveolar (pit) epithelium of the gastric mucosa. The transcript was expressed outwith the stomach only in metaplastic gastric epithelium: in the gut this was found in Barrett's oesophagus, duodenal mucosa or the ulcer-associated cell lineage (UACL); outwith the gut its presence was detected only in ovarian mucinous tumours. The mRNA was present in the stomach of other species, such as the mouse, in the same location as in humans.

During these studies, the transcript had also been identified by a Japanese research group who named it CA11. Since this name overlaps with that of carbonic anhydrase XI, to which the transcript bears no similarity, for the purposes of this thesis the gene has been re-named foveolin. The 5' and 3' ends of the mRNA have been characterised by Rapid Amplification of cDNA Ends (RACE). Homologous mouse and cow mRNAs have been identified, characterised and compared. The human gene has been partially cloned from a human genomic PAC library, and the full-length sequence has been obtained via the on-line Celera genomic databases and characterised. The PAC clone was used to map the gene by fluorescent *in situ* hybridisation (FISH) to human chromosome 2. Fragments of the mouse genomic sequence were identified in NCBI's on-line mouse genomic databases, assembled into the full-length sequence and characterised. The predicted protein product, like the mRNA and DNA sequences, is highly conserved between the human, mouse and cow species. The protein shows no homology to any known protein sequences or motifs, but bears an initial signal peptide and is therefore predicted to have an extracellular location, being either retained on the outer cell surface or secreted into the gastric lumen, much like gastric mucin (MUC5) and the trefoil peptide pS2 (TFF1), with which foveolin shares a similar location in the superficial and foveolar gastric epithelium. Except for the Japanese study, which presented only Northern blotting and RACE data for the human mRNA, this is the first detailed description of this new gene.

We believe that the results of our characterisation studies are valid. The new gastric gene, foveolin, is very highly expressed, mainly in normal stomach, and is absent from gastric carcinoma tissue and cell lines. These results agree with those of other researchers. Our Northern blotting (Figure 5.3 and Figure 5.4) and *in situ* hybridisation results (Figure 5.5 and Figure 5.6) are internally consistent and suggest that the mRNA is expressed only in native or metaplastic gastric epithelium. Only one other study of this gene has been published (Yoshikawa *et al.* 2000). Our results agree with almost of its data: the gene is expressed at high levels in normal stomach but is absent from gastric carcinomas and cell lines, and by Northern blotting, its mRNA is around 1 kb in size. The Japanese researchers found, as we did, that most tissues did not express the gene, including: small intestine, colon, liver, pancreas, trachea, lung, adrenal gland, kidney, prostate, testis, ovary, placenta, thyroid, heart, skeletal muscle, brain, spinal cord, peripheral blood lymphocytes, spleen, thymus, lymph node and bone marrow (Yoshikawa *et al.* 2000). Furthermore, most of the on-line cDNA libraries containing matching ESTs originate from stomach, whether from man or mouse.

However, the previous Japanese study and the site of origin of some of the ESTs in the on-line databases suggest that this gene may also be expressed outwith the gut, which conflicts with the results of our mRNA profiling by Northern blotting and *in situ* hybridisation. The Japanese study suggested that the gene is also expressed in uterus and placenta. Pregnant uterus was also a possible site of expression to judge by the on-line human ESTs, while matching mouse ESTs also originated from tongue, pancreas, small intestine, caecum and ovary/uterus (Wheeler *et al.* 2001).

By rapid amplification of cDNA ends (RACE), the Japanese group identified foveolin mRNA in samples from uterus and placenta (Yoshikawa *et al.* 2000). RACE is more commonly used (and was by us) only for characterisation of the 5'- and 3'-ends of cDNA. Because it is based on PCR, RACE is more sensitive than either Northern blotting or ISH, and would also be more sensitive to cross-contamination, perhaps especially when commercial sources of cDNA are used, as was done in this case. However, the results of the Japanese group were internally inconsistent: by Northern blotting, they found that placenta lacked detectable foveolin. Unfortunately, neither the Japanese group nor we have examined uterine tissue by Northern blotting.

Of the on-line human cDNA libraries containing foveolin sequences, the only ones which did not originate from the stomach were derived instead from pooled tissues including "pregnant uterus." Foveolin could therefore be expressed in uterus or in accompanying placenta. However, it could also be present in any foetal (gastric) tissue present but we could not obtain further details of the exact nature of these samples (personal communication, Marcelo Bento Soares, NCBI).

Of cDNA libraries containing foveolin in other species, the rat originated from a mixed-tissue pool and the cow originated only from abomasum, the fourth stomach. However, while most mouse ESTs were derived from stomach, a significant minority originated from other organs including tongue, pancreas, small intestine, caecum and ovary/uterus. Aside from the obvious possible explanation that these tissues do indeed express foveolin, another reason could be cross-contamination of these murine samples with gastric tissue, possibly during tissue dissection, since the mouse is relatively small.

In our studies, we found foveolin mRNA only in areas of native or metaplastic gastric epithelium. Expression outwith the gastro-intestinal tract was found only in ovarian mucinous tumours, which are well-recognised to show gastric and other forms of metaplasia. We examined a wide range of other tissues using ISH, including five sections of normal and abnormal uterine tissue, and including pancreas, small intestine and colon, but did not detect any foveolin.

These discrepancies mean that we cannot rule out the possibility that foveolin may not be truly gastric-specific, as we suggest, but may also be expressed at low levels in normal uterus and placenta, and possibly in other glandular tissues. It would therefore be appropriate to investigate the latter tissues further. Northern blotting of uterine tissue would be useful. RT-PCR or RACE could also be performed locally on appropriate samples. Lastly, if and when specific antibodies can be developed against foveolin, immunohistochemistry, and potentially Western blotting, would also be of use.

The phenomenon of gastric metaplasia in diverse tissues, including both inflammatory conditions of the gut and tumours of ovarian mucinous type, is intriguing. Metaplasia has been discussed many times already in this thesis (see Sections 1.3.1, 1.3.2 and 4.2.2, amongst others). Metaplasia in the stomach, from normal gastric epithelium towards an intestinal phenotype, is associated with loss of foveolin expression (Figure 5.6). Such intestinal metaplasia occurs in the stomach as a result of chronic gastritis, especially when associated with *H. pylori.* Conversely, gastric metaplasia is well-recognised to occur in the duodenum in response to epithelial damage and ulceration caused by excess acid, and this is associated with the acquisition of foveolin expression (Figure 5.6). Gastric metaplasia in the duodenum is thought to be but one form of the ulcer-associated cell lineage (UACL) (Kushima *et al.* 1999), a reparative phenomenon unique to the gastro-intestinal tract (Wong *et al.* 1999); its appearance elsewhere, for example in the colon, is also characterised by expression of foveolin.

Expression outwith the gut was found only in ovarian mucinous tumours. Malignant, borderline and benign ovarian mucinous tumours are all well-recognised to show gastric metaplasia of both superficial/foveolar and pyloric glandular types, in around

80% of cases (Tenti *et al.* 1992; Boman *et al.* 2001). Immunohistochemical and *in situ* hybridisation studies have shown that ovarian mucinous tumours often contain gastric mucin (MUC5) and/or pepsinogen C (Tenti *et al.* 1992; Boman *et al.* 2001). Since foveolin has a similar location to MUC5 in the gastric mucosa, its presence in these tumours is of no surprise. Out of interest, as well as showing gastric metaplasia, the mucin-secreting cells lining these ovarian tumours may also exhibit intestinal, pancreatico-biliary or endocervical phenotypes (Tenti *et al.* 1992; Boman *et al.* 2001). This phenomenon was discussed in the previous chapter in relation to both pancreatic and ovarian adenocarcinomas (Section 4.2.2), and indeed, it has previously been suggested that ovarian and pancreatic adenocarcinomas (especially mucinous cystic or serous subtypes) may share common origins (Zamboni *et al.* 1999). In our hands, however, foveolin was absent from the normal pancreas and from pancreatic adenocarcinomas.

This is the first study to characterise the detailed structure of the human, mouse and bovine mRNAs and their predicted proteins, plus the human and mouse genomic sequences. Comparison between species shows that the genes are highly conserved. The high degree of homology between the human, mouse and bovine mRNAs, proteins and genes supports our supposition that foveolin is likely to be functionally important in the normal stomach.

However, there are two possible translation start sites in the human foveolin mRNA which raises a debate about which of the two is more likely to be functional. Translation of proteins starts at an ATG codon (Strachan *et al.* 1999). By RACE, the Japanese group showed that the 5'-end of the human mRNA was around 50 bp longer than estimated by my RACE results, and 60 bp longer than the on-line ESTs (Yoshikawa *et al.* 2000). The extra 50 bp are consistent with the human genomic sequence which I later identified. This longer mRNA contains two such ATG codons, both in-frame with a long subsequent open reading frame. Normally translation is initiated at the upstream ATG: the "first ATG rule". This time, however, I believe, unlike the Japanese group, that the second ATG is likely to be the functional one, for a number of reasons.

First, alignment of the human, mouse and cow mRNA sequences shows that they are highly conserved around and downstream of the second human ATG, but the upstream segment shows much less homology. Second, this upstream difference is even more marked in the protein sequence. Although full upstream data is lacking for cow foveolin, comparison of its postulated translated sequence with that of the mouse and human shows that the 50 bp upstream segment is not at all similar. These mRNA and protein results can be seen by eye-balling the sequences and are confirmed by BLAST2 analysis. Third, all three species are predicted to contain a starting signal peptide, but this appears in the human only when the second ATG is used.

Fourth, start codons do not exist in isolation. For any given ATG to be functional, it must be surrounded by what is described as an appropriate nucleotide context (Kozak 1996). These desirable adjacent bases were first described by Marilyn Kozak and have hence been termed a Kozak sequence. The optimal context for initiation of translation in vertebrate mRNAs is regarded as **A**CCatg**G** (Kozak 1996). Within this consensus motif, nucleotides in two highly conserved positions exert the stongest effect: a G residue following the ATG codon (position +4) and a purine, preferably A, three nucleotides upstream (position –3) (Kozak 1996). However, this is not the whole story. The initiator context may vary according to whether the gene is located within sectors of DNA (isochores) which are GC-rich or GC-poor (Pesole *et al.* 1999). Genes located in GC-rich regions tend to require highly efficient translation and so are often house-keeping (constitutive, or to use the terminology of the previous chapter, "widely expressed") genes, whereas those located in GC-poor regions tend to require finer modulation and hence are usually tissue-specific (Pesole *et al.* 1999). Clearly, from our data, foveolin is likely to fall within the latter group. Pesole *et al* go on to list the ten most common initiation heptamers in genes from each of four groups of isochores of increasing GC-content.

For human foveolin, the context of the first possible site, which was favoured by the Japanese researchers, is **T**CCatg**C**. Kozak's preferred nucleotide is lacking from both of the key positions. None of the forty most common heptamers start with T and only two end with C, both in the groups with a higher GC-content (Pesole *et al.* 1999). The context of the second site is **A**AGatg**A**. Kozak's preferred nucleotide is fullfilled at the start, at position –3, although position +4 is not. Indeed, this heptamer is one of the ten most common for (tissue-specific) genes in GC-poor (i.e. tissue-specific) isochores (Pesole *et al.* 1999). The second ATG is thus obviously favoured, and in fact contradiction of the first ATG rule is common in tissue-specific genes (60%) compared to house-keeping genes (30%) (Pesole *et al.* 1999).

The initiator context of the one and only ATG in the cow mRNA sequence is identical to that in the human and therefore favourable. The mouse mRNA sequence starts directly with the first ATG so its context cannot be fully assessed. However, the nucleotides at the second site are **G**CCatg**A**. Again, Kozak's preferred purine nucleotide is fullfilled at the start, at position –3, although position +4 is not; and again, this heptamer is one of the ten most common for (tissue-specific) genes in relatively GC-poor isochores (Pesole *et al.* 1999). Thus, the nucleotides immediately surrounding the second ATG in the human mRNA sequence constitute a good initiator context, which is also the case for the equivalent ATGs in the mouse and cow sequences, whereas the ATG 50 bp upstream in the human sequence definitely lacks a favourable context.

Assuming that the second ATG is functional, then the predicted protein has a signal peptide. A signal peptide comprises the first 20 or so amino acids at the N-terminal end of a protein, always includes a substantial proportion of hydrophobic amino acids (Strachan *et al.* 1999) and is present in around 20% of human genes (Lander *et al.* 2001). The signal peptide enables the protein to be guided from its site of synthesis on the cytoplasmic ribosomes to the lumen of the endoplasmic reticulum and thence for secretion, at which point the signal peptide is cleaved from the mature protein. The presence of such a signal peptide in the predicted foveolin protein thus means it is likely to be extracellular, being either retained on the outer cell surface or secreted into the gastric lumen, much like gastric mucin (MUC5) and the trefoil peptide pS2 (TFF1) (Wong *et al.* 1999; Corfield *et al.* 2000).

The predicted foveolin protein has a length of around 185 amino acids in each species and a molecular weight of just over 20 kDa; these calculated figures include the signal peptide, which would be absent from the mature protein. For comparison, pS2 also bears a signal peptide and is secreted; the mature protein is 60 amino acids long with a molecular weight of 6.5 kDa (Jakowlew *et al.* 1984; Newton *et al.* 2000). Mucins, in contrast, are complex molecules, of high molecular weight, with extensive post-translational modification through the addition of carbohydrate moieties (Corfield *et al.* 2000).

Foveolin is very highly expressed in normal stomach and is therefore likely to be important in its function; its exact role, however, remains unclear. In normal human gastric antrum, to judge by SAGE, foveolin was one of the ten most highly expressed genes: it represented 0.8% of mRNA. Its abundance has been even higher, up to 5.7%, in normal mouse gastric cDNA libraries. Better characterised genes which have similar expression levels in normal stomach and which are likewise down-regulated in gastric carcinoma include gastrin, lysozyme, MUC5 and pS2 (see Section 4.2.2). These genes are all important in gastric function, providing supporting evidence for a similar role for foveolin.

Foveolin is located in the superficial/foveolar epithelium of the gastric mucosa. Its site in the stomach is similar but not identical to MUC5 and pS2 (see Figure 5.5) (Machado *et al.* 2000). MUC5 and pS2 tend to be co-localised in glandular mucosae, and play a role in mucosal protection, lubrication and repair (reviewed in (Williams *et al.* 1997; Wong *et al.* 1999; Corfield *et al.* 2000)). Outwith the superficial/foveolar epithelium of the gastric mucosa, MUC5 and pS2 are also co-expressed elsewhere in the normal and diseased gastro-intestinal tract including: the normal small intestine, colon and gall bladder; peptic and other ulcers, in the ulcer-associated cell lineage (UACL); intestinal metaplasia of the stomach and gastric metaplasia of Barrett's oesophagus (Wong *et al.* 1999; Corfield *et al.* 2000). Beyond the gut, these molecules are expressed in normal respiratory tract and uterus, where MUC5 is the main secretory mucin, and in adenocarcinomas of the breast, stomach, pancreas, ovary and uterus amongst others (Henry *et al.* 1991). Because of its similar location to MUC5 and pS2 in the stomach, it was at first thought that foveolin might also be present in these gynaecological and glandular tissues. However, as discussed above, our study, unlike that of the Japanese group (Yoshikawa *et al.* 2000), found no foveolin expression in any of these tissues, except in the well-recognised condition of gastric metaplasia.

The lack of any homology with any known proteins or recognised protein motifs by database searching hampers our efforts to propose the likely function of foveolin. However, although no structural evidence exists, one role in the normal stomach which could be postulated is that of protection against acid damage. I have tried and failed to identify any proteins known to have such a function to enable comparison with foveolin, so this remains entirely speculative.

Is down-regulation of foveolin in gastric carcinoma likely to be functionally important in tumorigenesis? Again, pS2 is used as an analogy. pS2 is expressed in most adenocarcinomas, including those of stomach, breast, pancreas, lung, uterus and ovary (Henry *et al.* 1991). In these tumours, the gene appears largely intact at the DNA level (Luqmani *et al.* 1989; Williams *et al.* 1997), but its expression at the mRNA and protein levels tends to be down-regulated in comparison with normal (Henry *et al.* 1991). This hold true for gastric carcinomas (Luqmani *et al.* 1989; Machado *et al.* 1996; Machado *et al.* 2000). Moreover, in a knockout mouse model, homozygous pS2-null mice develop hyperplasia, dysplasia and carcinoma of the gastric antral mucosa, which suggests that pS2 is important in normal gastric function and may act as a gastric-specific tumour suppressor gene (Lefebvre *et al.* 1996). Purely on the basis of their similar expression patterns, a similar role could be postulated for foveolin.

However, foveolin could be located precisely on human chromosome 2 through the presence of a sequence tagged site (D2S292-D2S145) in the mRNA sequence and through FISH (Figure 5.9). Genetic instability at this locus has not been reported.in gastric carcinoma (Mitelman *et al.* 1997; Yoshikawa *et al.* 2000).

Characterisation and functional analysis of the foveolin protein is the next step. Characterisation and expression profiling of the mRNAs, investigation of the genomic sequence, and inter-species comparisons of foveolin are important but form only part of its analysis. The next steps would be to characterise foveolin further at the protein level and to investigate its likely function through its expression in cells *in vitro* and through *in vivo* experiments.

For protein characterisation, specific antibodies are required. These would be generated as rabbit polyclonals, using synthetic small peptides and full-size foveolin protein as immunogens. The latter would be generated by cloning full-length foveolin cDNA into a multiple cloning expression system. The cDNA, fused to a molecular tag, would be expressed *in vitro* in cells and, following large-scale production, the full-size foveolin protein would be retrieved, via its tag. The antibodies would be validated on protein extracts of normal stomach then used to characterise the foveolin protein and to profile its expression in various tissues by Western blotting and by immunohistochemistry. These results would be expected to be broadly similar to, and to confirm those from, Northern blotting and *in situ* hybridisation. In addition, however, Western blotting would provide an indication of the size of the mature foveolin protein, and immunohistochemistry would confirm whether the protein is secreted from the foveolar epithelium into the gastric lumen.

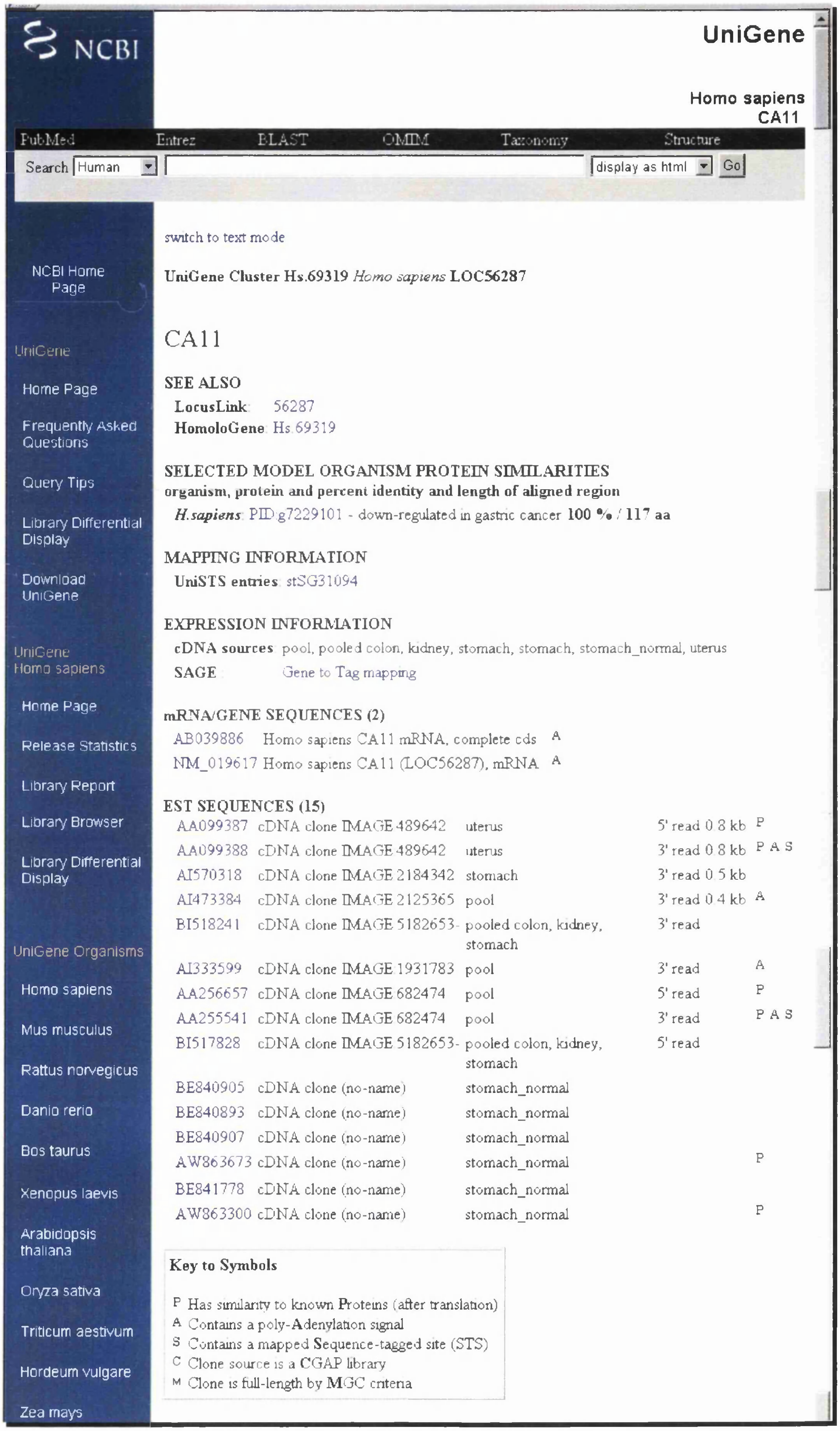
The sub-cellular location of foveolin would be studied by producing it in cells *in vitro* fused to the marker green fluorescent protein (GFP). The fusion protein would then be tracked by confocal microscopy. In order to investigate its biological function *in vitro*, foveolin would be expressed in gastric carcinoma cell lines, from which it is normally absent, to find out whether this alters cell morphology, or cell behaviour by invasion assays. Lastly, *in vivo* functional assays would be performed. Using the foveolin genomic sequences already generated, foveolin would be over-expressed, in transgenic mice, or inactivated, through mouse knock-out models. The phenotypic and functional consequences would be investigated, to check whether, for example, the gastric mucosa was morphologically different and whether acid levels were altered. The effect of *H. pylori* infection would also be studied. Based on its similar expression pattern, we have postulated that foveolin could, like pS2 (Lefebvre *et al.* 1996), play a role as a gastric-specific tumour suppressor gene. This hypothesis can only be investigated through the use of animal models, since such functional effects cannot be demonstrated by *in vitro* work alone.

Conclusion. Large-scale mRNA expression profiling of gastric adenocarcinoma and normal stomach by SAGE produced a number of candidate molecules. One was a novel gene which was highly and specifically expressed in normal stomach, at levels similar to gastrin, MUC5 and the trefoil peptides, which are important in normal gastric physiology. The gene is expressed in native or metaplastic gastric surface and foveolar epithelium and hence has been named foveolin. Although it appears to be gastric-predominant, foveolin may also be expressed at low levels in the uterus, although evidence for this is conflicting. Foveolin is not expressed in gastric adenocarcinomas or cell lines.

Foveolin mRNAs and their predicted proteins have been structurally characterised from human, mouse and cow, as have the human and mouse genomic sequences. The mRNAs are around 700-800 bp in length. The genes contain six exons which extend over around 6 kb. In the human, the gene maps by FISH to chromosome 2. The proteins are around 186 amino acids in length and just over 20 kDa in molecular weight. The sequences are highly homologous between species, again suggesting that it plays an important role in the stomach. Foveolin lacks homology to any known protein sequence or motif, so its function can only be postulated on the basis of its mucosal location: it could involve mucosal protection and possibly defence against gastric acid. This is supported by the fact that foveolin's primary sequence contains a signal peptide, indicating that the protein is extracellular, either located on the cell surface or secreted. Its absence from gastric carcinomas suggests that it may act as a functional gastric tumour suppressor gene, analogous to pS2 (trefoil factor 1): pS2 is highly expressed in normal stomach and down-regulated in tumours and pS2 knock-out mice develop gastric dysplasia then carcinoma. Further characterisation of foveolin's biological role is therefore required and likely to be worthwhile.

Figure 5.1: UniGene cluster for the new gastric gene

When the short SAGE tag of interest was entered into SAGEmap, it matched a single UniGene cluster. At the time, this cluster contained only eight uncharacterised cDNAs, known as expressed sequence tags (ESTs). However, it now contains 2 mRNA/gene sequences and 15 EST sequences, which are used to provide the "Expression Information". The mapping information relates to a single sequence tagged site (STS) present in the mRNA sequence. The eight matching ESTs were down-loaded and further analysed.



UniGene Cluster Hs.69319 *Homo sapiens* **LOC56287**

CA11

SEE ALSO

LocusLink: 56287
HomoloGene: Hs.69319

SELECTED MODEL ORGANISM PROTEIN SIMILARITIES
organism, protein and percent identity and length of aligned region

H.sapiens: PID:g7229101 - down-regulated in gastric cancer **100 % / 117 aa**

MAPPING INFORMATION

UniSTS entries: stSG31094

EXPRESSION INFORMATION

cDNA sources: pool, pooled colon, kidney, stomach, stomach, stomach_normal, uterus
SAGE: Gene to Tag mapping

mRNA/GENE SEQUENCES (2)

AB039886	Homo sapiens CA11 mRNA, complete cds	A
NM_019617	Homo sapiens CA11 (LOC56287), mRNA	A

EST SEQUENCES (15)

AA099387	cDNA clone IMAGE:489642	uterus	5' read 0.8 kb	P
AA099388	cDNA clone IMAGE:489642	uterus	3' read 0.8 kb	P A S
AI570318	cDNA clone IMAGE:2184342	stomach	3' read 0.5 kb	
AI473384	cDNA clone IMAGE:2125365	pool	3' read 0.4 kb	A
BI518241	cDNA clone IMAGE:5182653-	pooled colon, kidney, stomach	3' read	
AI333599	cDNA clone IMAGE:1931783	pool	3' read	A
AA256657	cDNA clone IMAGE:682474	pool	5' read	P
AA255541	cDNA clone IMAGE:682474	pool	3' read	P A S
BI517828	cDNA clone IMAGE:5182653-	pooled colon, kidney, stomach	5' read	
BE840905	cDNA clone (no-name)	stomach_normal		
BE840893	cDNA clone (no-name)	stomach_normal		
BE840907	cDNA clone (no-name)	stomach_normal		
AW863673	cDNA clone (no-name)	stomach_normal		P
BE841778	cDNA clone (no-name)	stomach_normal		
AW863300	cDNA clone (no-name)	stomach_normal		P

Key to Symbols

P	Has similarity to known Proteins (after translation)
A	Contains a poly-Adenylation signal
S	Contains a mapped Sequence-tagged site (STS)
C	Clone source is a CGAP library
M	Clone is full-length by MGC criteria

Figure 5.2: Schematic diagram of the eight ESTs in the UniGene cluster

The eight ESTs in the matching UniGene cluster were down-loaded and assembled using SEQMAN. It became obvious that the ESTs overlapped and they were merged to create a single cDNA sequence with a length of around 750 base-pairs. Further analysis showed that the cDNA sequence contains a long open reading frame (ORF, or translatable protein sequence), suggesting this is a true gene. The cDNA fragment was amplified by polymerase chain reaction, cloned and used as a probe for Northern blotting and *in situ* hybridisation.

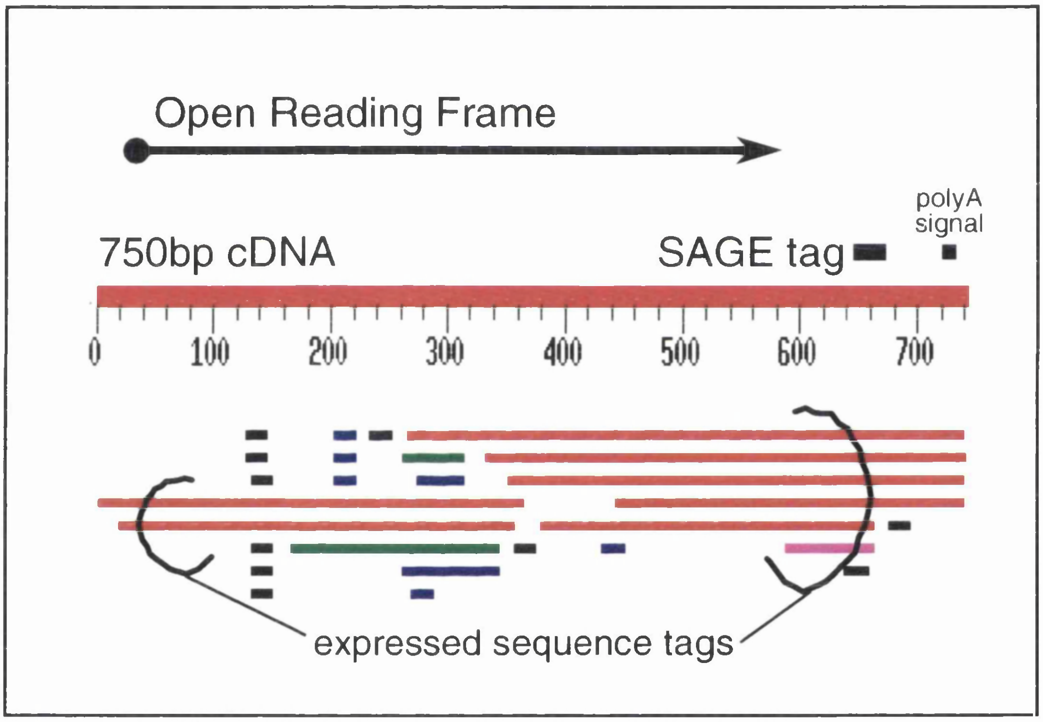

Open Reading Frame
polyA
signal
750bp cDNA
SAGE tag
0
100
200
300
400
500
600
700
expressed sequence tags

Figure 5.3: Northern Blot for the new gastric gene with gastro-intestinal samples

The same Northern blot was used as in the previous chapter: RNA was isolated from 19 gastro-intestinal tumor and normal tissues and cell lines, as indicated along the top row, of which further details are listed in Table 4.3. Hybridisation with a radio-labelled probe for foveolin mRNA showed its presence only in, but at high levels in, normal stomach, in all of its anatomical areas, namely the cardia, body and antrum. Foveolin was absent from all of the other gastro-intestinal samples, including: gastric adenocarcinomas and gastro-oesophageal adenocarcinoma cell lines; normal squamous oesophagus, small intestine and colon, and their corresponding tumours, oesophageal squamous carcinoma and colonic adenocarcinoma.

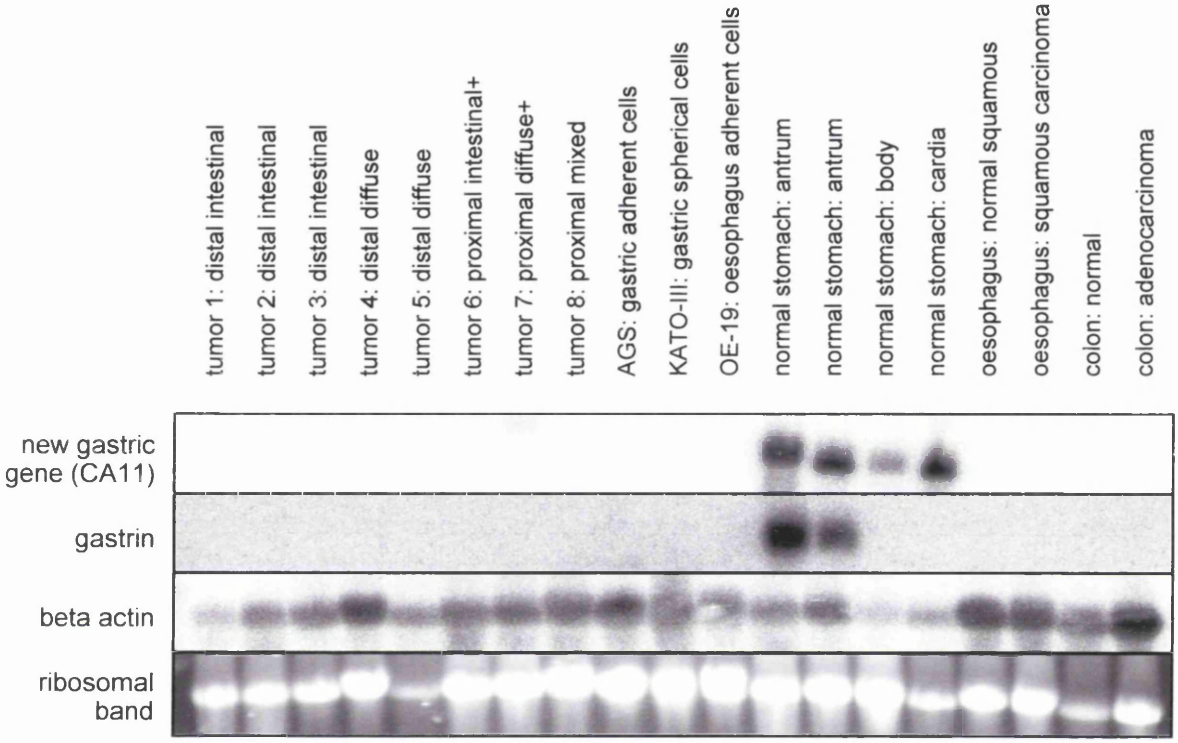

tumor 1: distal intestinal
tumor 2: distal intestinal
tumor 3: distal intestinal
tumor 4: distal diffuse
tumor 5: distal diffuse
tumor 6: proximal intestinal+
tumor 7: proximal diffuse+
tumor 8: proximal mixed
AGS: gastric adherent cells
KATO-III: gastric spherical cells
OE-19: oesophagus adherent cells
normal stomach: antrum
normal stomach: antrum
normal stomach: body
normal stomach: cardia
oesophagus: normal squamous
oesophagus: squamous carcinoma
colon: normal
colon: adenocarcinoma
new gastric gene (CA11)
gastrin
beta actin
ribosomal band

Figure 5.4: Commercial Northern blots for the new gastric gene with a wide range of tissues

The probe was then applied to two more, commercial, Northern blots: Invitrogen's Northern Territory^{TM} Human Normal Tissue Blot III, and OriGene Technologies' Multiple Choice^{TM} Northern Blot. The tissues from which the total RNA originated are indicated along the top row.

The lower row shows that, again, foveolin was expressed only in normal stomach. Foveolin mRNA was again absent from other gastro-intestinal tissues, including colon, appendix, gallbladder and liver. It was also absent from lymphoid organs, including tonsil, thymus, lymph node and spleen, and from genito-urinary organs including prostate, testis, ovary and placenta.

In the upper row, hybridisation with a loading control probe for β-actin confirms the presence of good quantities of intact RNA in each sample, confirming that foveolin mRNA expression is truly absent.

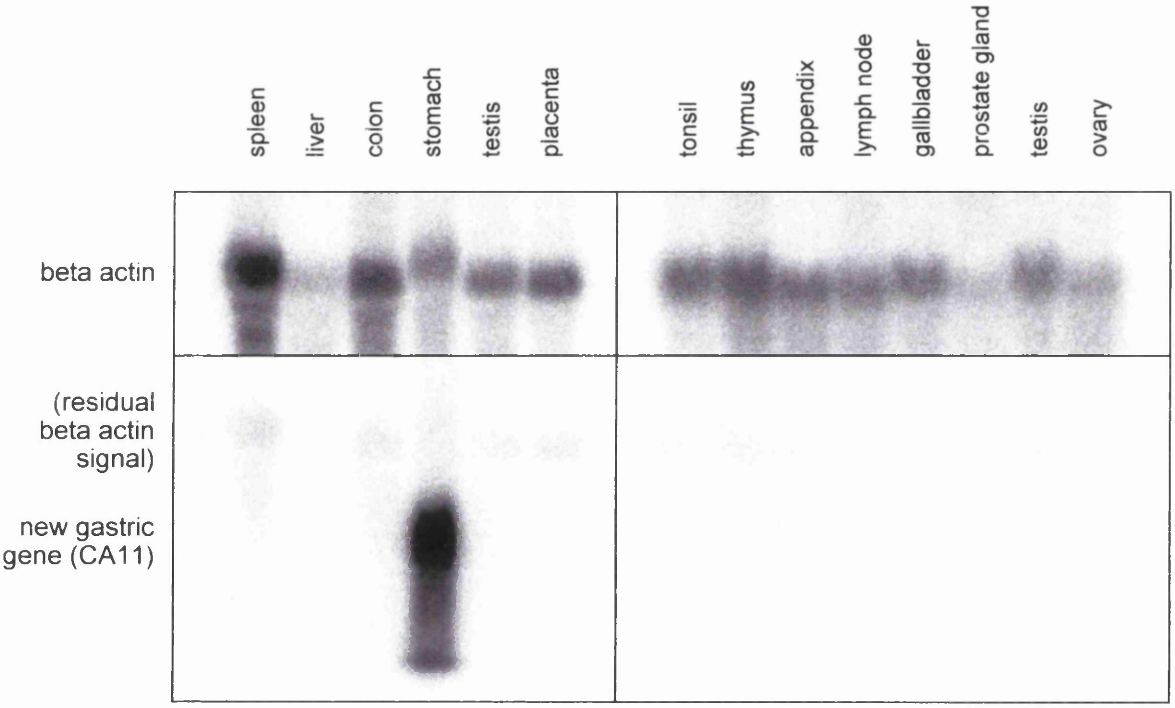

spleen
liver
colon
stomach
testis
placenta
tonsil
thymus
appendix
lymph node
gallbladder
prostate gland
testis
ovary
beta actin
(residual beta actin signal)
new gastric gene (CA11)

Figure 5.5: *In situ* hybridisation for the new gastric gene in normal stomach

a. In the positive control slides, mRNA for the new gene, foveolin, was localised to the superficial/foveolar zone of normal gastric gastric mucosa (see Section 1.2.2). The positive ISH staining is coloured blue-black. In normal gastric body mucosa, the superficial/foveolar epithelium occupies the upper quarter of the gastric glands. The deeper parts of the gastric glands comprise the proliferative zone, which occupies the next quarter of the glands, and the gastric pits, which make up the lower half of the glands, and these did not contain foveolin.

b. This negative control contains normal gastric body subjected to ISH but with the antisense probe omitted. Staining for foveolin mRNA is absent. Hybridisation with a further negative control, the sense probe, yields an identical appearance.

c. This gastric mucosa shows double-staining for foveolin (black) and pS2 (brown). Foveolin mRNA is once more stained blue-black by ISH, which was followed by (brown) immunohistochemical staining for pS2 (trefoil factor 1 (TFF1)). Like foveolin, pS2 is present in the superficial epithelium; but unlike foveolin, pS2 is also found in the deeper gastric glands and in the goblet cells of intestinal metaplasia (arrow).

d. Normal mouse gastric mucosa contains foveolin mRNA in the superficial/foveolar epithelium, identical to the human location.

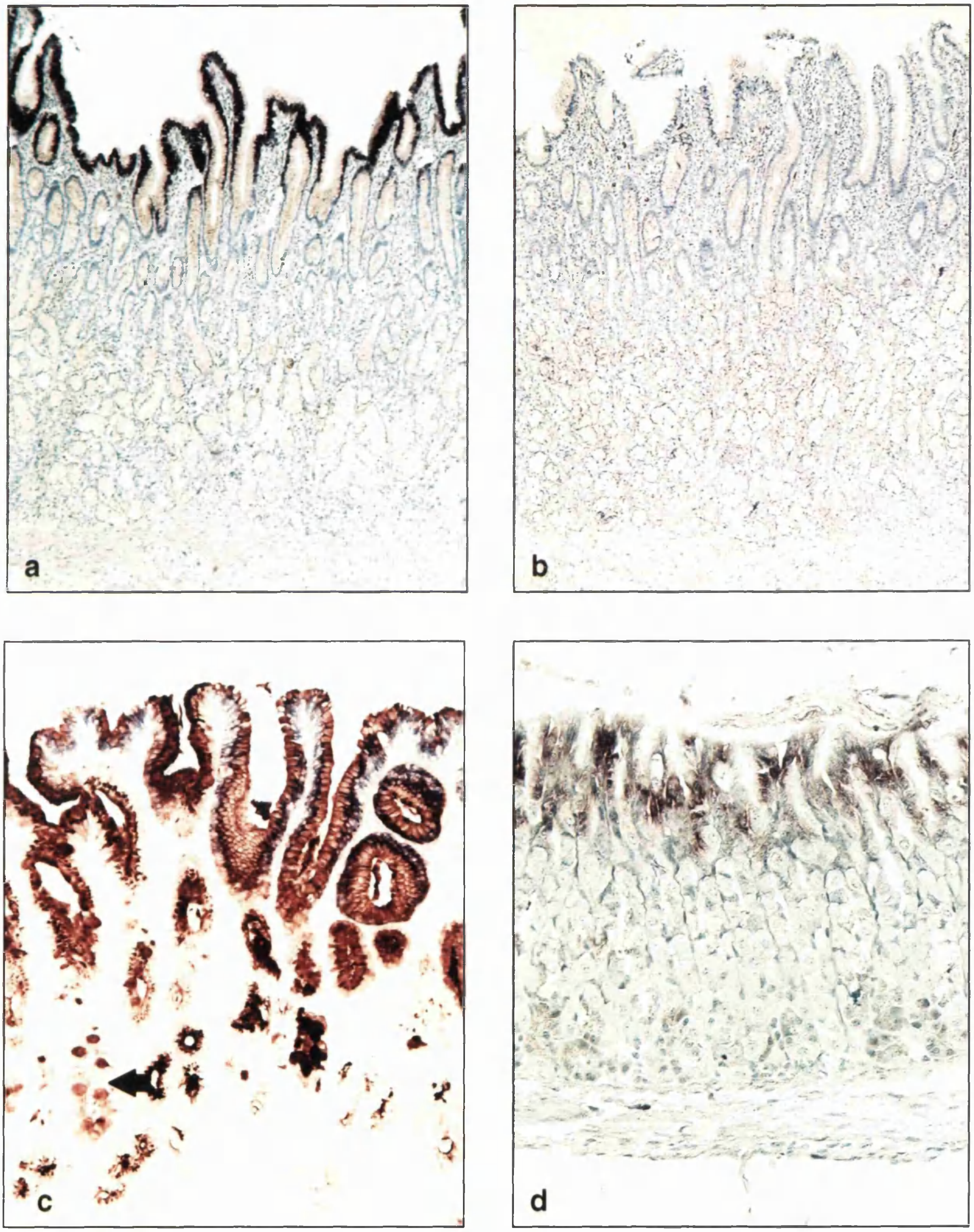

a
b
c
d

Figure 5.6: *In situ* hybridisation for foveolin in other normal and diseased upper gastro-intestinal tissues

a. Metaplastic epithelium in the stomach is usually of intestinal type and contains goblet cells which stain positively with the turquoise Alcian Blue stain for acidic mucins (see Sections 1.3.1 and 1.4.6). Foveolin mRNA staining is absent from intestinal metaplasia but remains in the adjacent remaining normal gastric epithelium. The change between the two types of epithelia is abrupt both in terms of foveolin ISH staining and histological appearance.

b. Unlike the adjacent normal gastric mucosa, dysplastic (cancerous) gastric epithelial cells cells (arrow) do not contain foveolin mRNA.

c. Foveolin mRNA is absent from normal duodenal epithelium which is of intestinal type, contains goblet cells and stains positively with the turquoise Alcian Blue. Blue-black staining for foveolin is present only in small (abnormal) foci of gastric metaplasia at the tips of the small intestinal villi.

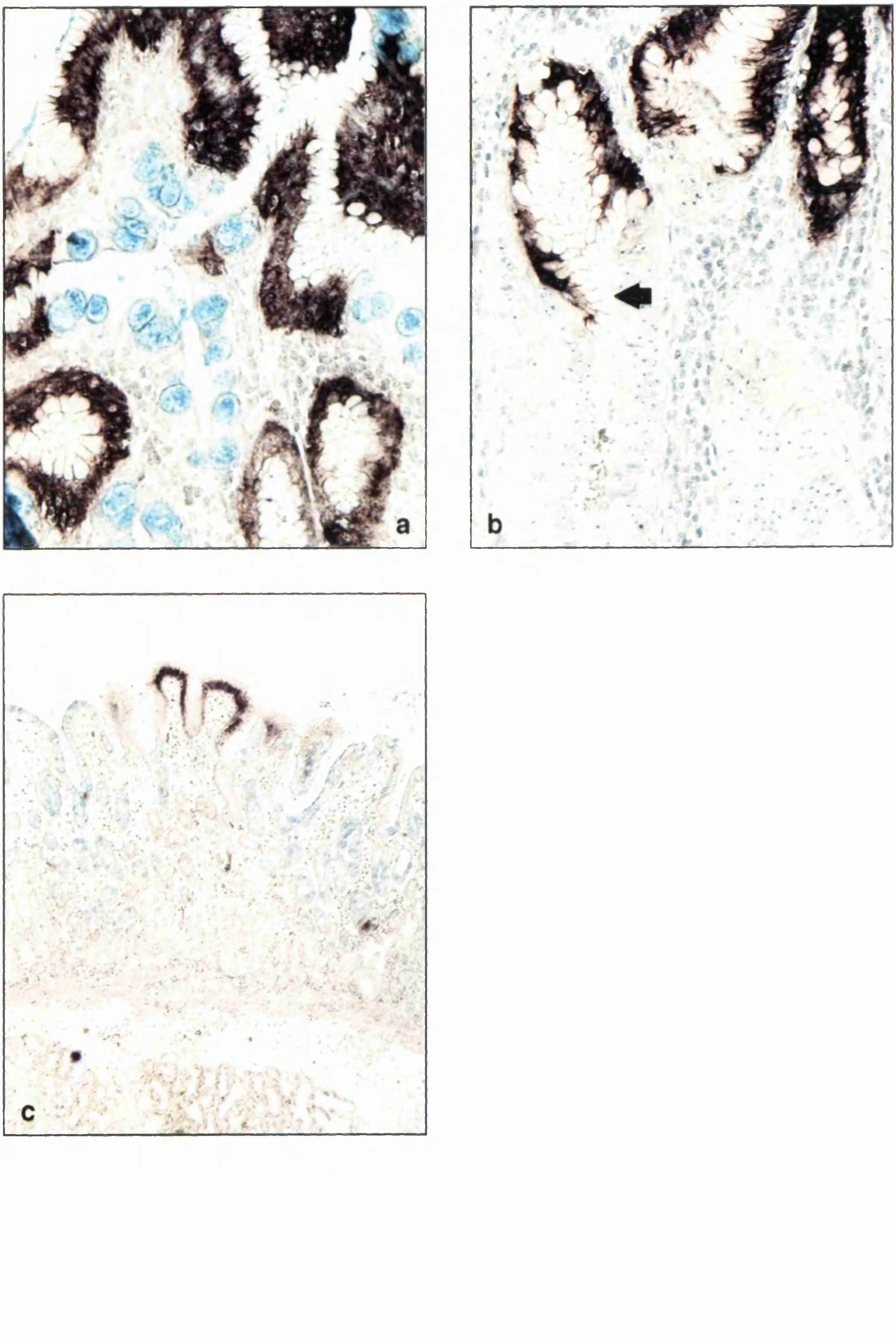

a
b
c

Figure 5.7: Foveolin cDNA sequences in human, mouse and cow

Here, the cDNA sequences for human, mouse and cow are presented and compared. These are composite results derived from my 3' and 5' RACE, the Japanese 3' and 5' RACE and the on-line sequences. Similar nucleotides are highlighted: those identical across all three sequences have a black background, whereas nucleotides shared by two sequences have a grey background.

Near the start of the mRNAs, the upstream (5') ATGs in the human and mouse sequences are indicated in bold italic type. For the mouse, my 5' RACE sequence starts at position 60 (GAGCAC...), but the immediately upstream 8 bp of genomic sequence contain another ATG, and are included here for the purposes of debate. This mouse mRNA is longer than any of the many on-line ESTs, but my other 5' RACE sequence, for human foveolin, proved to be 50 bp short so it is possible that the same could apply here.

The ATG which is more likely to be functional (the second in the human and mouse sequences) (see Discussion Section 5.2) is indicated in bold type.

The exon boundaries, identified by comparison with the genomic sequnces (see Sections 5.1.2.4 and 5.1.2.5, and Figure 5.10 and Figure 5.12) are indicated by arrows.

The polyadenylation site AATAAA is indicated in bold type.

```
Human  ATAACACCTAGTTTGAGTCAACCTGGTTAAGTACAAATATGAGAAGGCTT 50
Mouse  --------------------------------------------------
Cow    --------------------------------------------------

        upstream    upstream
        mouse       human
        ATG         ATG
Human  CTCATTCAGGTCCATGCTTGCCTACTCCTCTGTCCACTGCTTTCGTGAAG 100
Mouse  -ATGTCTGTGAGCACGCCGCCTGATCCTCTGCTCCACCACACTCCTGCCG
Cow    ----------------------TCATCTCAGCTTCATTGCTTTGGTGAAG

            probable
            functional
            ATG        ↓
Human  ACAAGATGAAGTTCACAATTGTCTTTGCTGGACTTCTTGGAGTCTTTCTA 150
Mouse  CC---ATGAAGCTCACAATGTTCGTCGTGGGTCTGCTTGGCCTCCTTGCA
Cow    TCAAGATGAAGTTCACAATTGTCTTTGCTGGACTTCTTGGCATCTTCCTG

                           ↓
Human  GCTCCTGCCCTTGCTAACTATAATATCAACGTCAATGATGACAACAACAA 200
Mouse  GCTCCTGGTTTTGCT---TACACGGTCAACATCAATGGTAATGATGGCAA
Cow    ACTCCTACTCTTGCTGACTATGATATCAGTGTTAATGATAACAACAACAG

Human  TGC---TGGAAGTGGGCAGCAGTCAGTGAGTGTCAACAATGAACACAATG 250
Mouse  TGTAGACGGAAGTGGACAGCAGTCGGTGAGCATCAATGGTGTGCACAACG
Cow    TGG---TGGAAGTGGGCAGCAGTCAGTGAGTGTCAACAATGAACATGGCG

Human  TGGCCAATGTTGACAATAACAACGGATGGGACTCCTGGAATTCCATCTGG 300
Mouse  TGGCCAATATCGACAACAATAACGGCTGGGACTCCTGGAATAGCCTCTGG
Cow    TGGCCAATGTTGACAATAACAATGGATGGGACTCCTGGAACTCCCTCTGG

                  ↓
Human  GATTATGGAAATGGCTTTGCTGCAACCAGACTCTTTCAAAAGAAGACATG 350
Mouse  GACTATGAAAACAGTTTCGCTGCCACGAGACTCTTCTCCAAGAAGTCATG
Cow    GATTACGGAAGTGGCTTTGCTGTCATCAGACCCTTTAAGAAGAAGTCGTG

Human  CATTGTGCACAAAATGAACAAGGAAGTCATGCCCTCCATTCAATCCCTTG 400
Mouse  CATTGTGCACAGAATGAACAAGGATGCCATGCCCTCCATTCAGGACCTCG
Cow    CATTGTGCACAAAATGAACAAGGAAGTCATGCCCTCTATTCAAGCCCTTG
```

```
Human  ATGCACTGGTCAAGGAAAAGAAGCTTCAGGGTAAGGGACCAGGAGGACCA 450
Mouse  ATACAATGGTCAAGGAACAGAAG------GGTAAAGGGCCTGGAGGAGCT
Cow    ATATGCTGGCCAAGAAAAACAAGCTTCAGGGTAGAGGACCAGAGGGACCA

Human  CCTCCCAAGGGCCTGATGTACTCAGTCAACCCAAACAAAGTCGATGACCT 500
Mouse  CCTCCCAAGGACTTGATGTACTCCGTCAACCCTACCAGAGTGGAGGACCT
Cow    CCTCCCAAGAGCCTGATCTACTCAGTCAAGCCTGACAAAGTCAACAACCT

Human  GAGCAAGTTCGGAAAAAACATTGCAAACATGTGTCGTGGGATTCCAACAT 550
Mouse  GAATACATTCGGACCAAAGATTGCTGGCATGTGCAGGGGCATCCCTACCT
Cow    GGACCAGTTTGGAAAATTCATTGTTACCATGTGCAAGGGGATTCCAACAT

Human  ACATGGCTGAGGAGATGCAAGAGGCAAGCCTGTTTTTTTACTCAGGAACG 600
Mouse  ATGTGGCCGAGGAGATTCCAGGACCAAACCAGCCTTTGTACTCAAAGAAG
Cow    ACATGGCTGAAGAGATTCAGGGAGCAAACCTGATTCTGTACCCAGAAAAG

Human  TGCTACACGACCAGTGTACTATGGATTGTGGACATTTCCTTCTGTGGAGA 650
Mouse  TGCTACACAGCTGACATACTCTGGATTCTGCGGATGTCCTTCTGTGGAAC
Cow    TGCTTCAATGTTGATATACTCTGGATTCTGAACATTTCCTTATGTGAAGA

Human  CACGGTGGAGAACTAAACAATTTTTTAAAGCCACTATGGATTTAGTCGTC 700
Mouse  ATCAGTGGAGACATA-----------------------------------C
Cow    AGCAATGGAGAACTAA----------------------------------

Human  TGAATATGCTGTGCAGAAAAAATATGGGCTCCAGTGGTTTTTACCATGTC 750
Mouse  TAGAAGTCACAGGAAAACAACCCGTGGGCTCTGACCATCGCAATGCT-TG
Cow    --------------------------------------------------

                                                    AAT
Human  ATTCTGAAATTTTTCTCTACTAGTTATGTTTGATTTCTTTAAGTTTCAAT 800
Mouse  ATTATGAGAGTGTTCTCTGGGGGTTGTGATTAGCTTCTTTAAGGCTCAAT
Cow    --------------------------------------------------

       AAA
Human  AAAATCATTTAGCATTG-- 819
Mouse  AAACCCACGTGGCAGCACA
Cow    -----------------
```

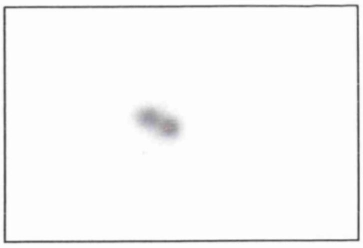
Figure 5.8: Identification of a human foveolin clone through probing a genomic library

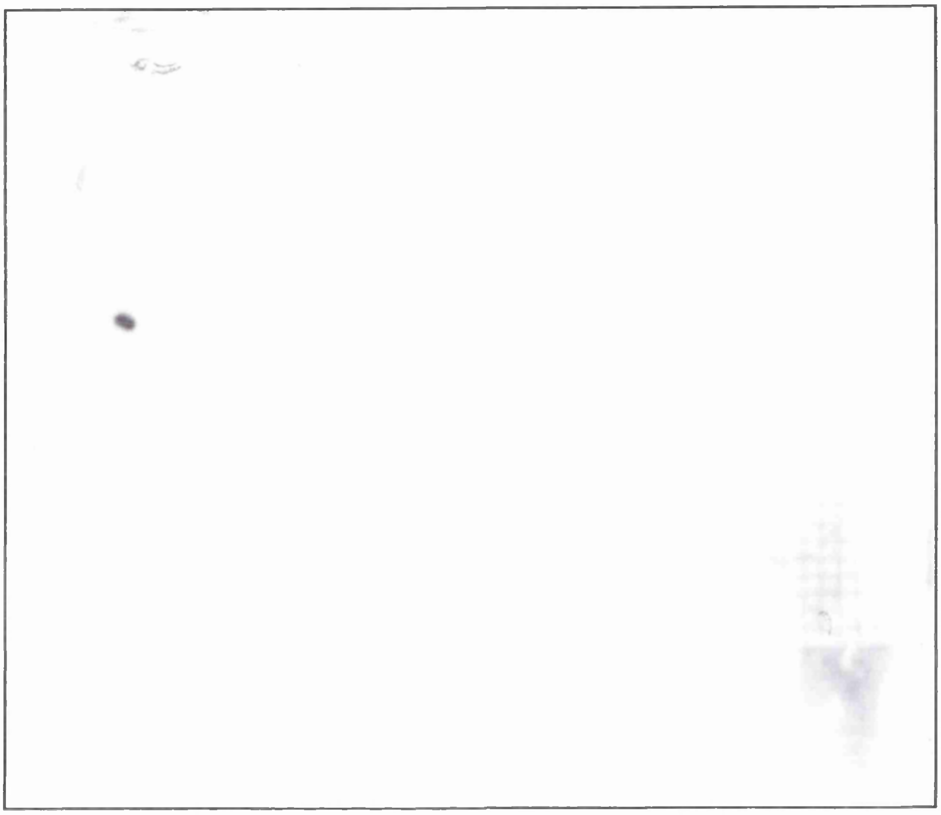
This figure illustrates the seventh high-density gridded filter of the genomic PAC library RPCI1, containing the single positive human foveolin genomic clone.

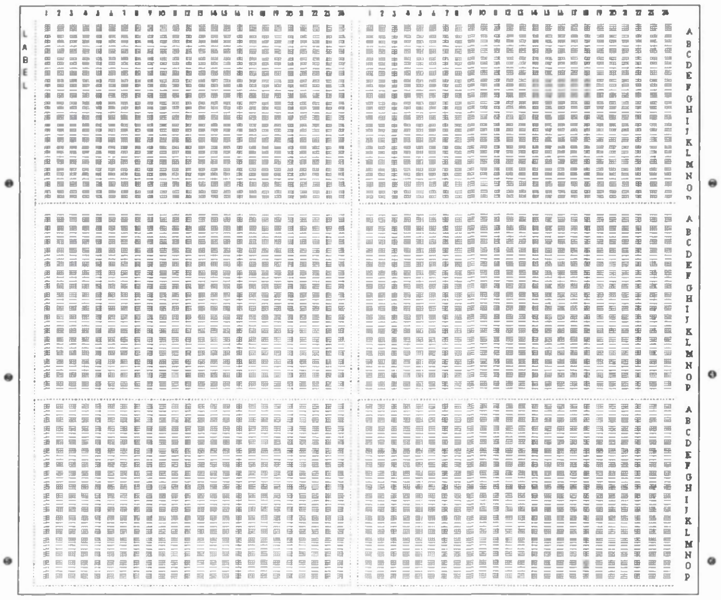
The human PAC library RPCI1 is a genomic library within PACs (plasmid artificial chromosomes) with an average insert size of about 110 kb (Osoegawa *et al.* 2001). The library is spotted on seven high-density gridded filters and was obtained from HGMP. The previously described radiolabelled cDNA probe was used to probe the PAC libraries. Unfortunately, some of the blots had been hybridised before and blots 1-4 were found to have been irrevocably damaged by stripping. Luckily, the remaining intact blots (5-7) contained a single positive clone in blot 7, as shown here.

At the bottom is a diagram explaining the orientation of the filter. Above is a low-power view of the scanned filter hybridised with the foveolin probe. At the top is a higher-power view of the single spot. Identification of the positive clone can be made from its position in the grid on the filter and from the orientation of the two positive spots within the 16-spotted square.

This clone was obtained and a DNA plasmid preparation was performed. The cloned PAC DNA was then used as a probe for fluorescent *in situ* hybridisation (FISH) and also for genomic sequencing.







1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
LABEL
A B C D E F G H I J K L M N O P
A B C D E F G H I J K L M N O P
A B C D E F G H I J K L M N O P

Figure 5.9: Fluorescent *in situ* hybridisation (FISH) for the human foveolin gene

This figure illustrates the results of fluorescent *in situ* hybridisation (FISH). The human foveolin gene can be seen with the green colour. Because this technique uses a metaphase spread of chromosomes which have already duplicated but not separated, four copies of the gene can be visualised. The UniGene cluster for the human gene describes a matching sequence tagged site (STS) (UniSTS entry: stSG31094) on Chromosome 2. The pink colour is so-called paint for chromosome 2 which confirms this as the gene's location, on the small arm close to the centromere, in agreement with the STS prediction.

Figure 5.10: The full sequence of the human foveolin gene

This figure shows the human genomic sequence of foveolin. This, the full sequence, was obtained by interrogating the Celera web-site, although its 3' end (beyond position 5194, indicated with an arrow) had already been obtained using the PAC library and sequencing of the single resulting clone. The exons were identified by comparison with the known mRNA sequences and are highlighted in yellow.

TATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCATGGAACATCTTCCAGACCAGCATTTCACAGGTCCACTGC 100
AGAACTTCAGGCCCTTGTCCTTGGTATGACATGCAAAATGTGTGCTGGGCACTGCACCTGTGTACCAAGATGACTACTGGTGGCTGAGTTTAAACGTAGA 200
GACCATGCTACCCACTCATTGACTTCTCAGTGACATGAAGTTTCCTAAAGCAGATGAAGGCCTCTTCACACATGTCTGCCCCTTACTCCAGGCTTATTAC 300
CTTCTGTGGATAAAGAGTGAGTTTGAAAGAAATCAGTGGACACATTAAAAAATATCCAGTAATACTGCAGCTGTTCTAATCCACTCCTACCAACAGTTTT 400
GAAGAGATGTTGATAGCGAATCAAGCTTTATAACCATGTGATCCCATCTTATGGTTTCAATCCATGCACAGGAGGAAAATTGTGGGCACGAAGTTTCCAA 500
AGGGAAAATTTATAGATTGGTAGTTAATGAAATACAGTTTTCCTCCTTGGCAAATTTAATTTACTAGCTTCACTGTATAGGAAAAAGCAGGAAAAAAATT 600
AAAACCAACTCACCTCCAAACCTGTTTTGAGCTTTTACTTGTCTGCCCAATTGATAGTTTCTACTCTCTGCTTTTGATGAAAATATTTTTTATTATTTTA 700
ATGTAACTTCTGAAAACTAAATTATCTAGAAGCAAATAAAAAGATATTGCTTTTATAGTTCCCAGAAGGAAAAAACAAACACTAGGAAAGTTCTATCTAT 800
CAGATGGGGGAGATGTGATGGAGGCAGTGATATTTGAGCTGAGCCTTGAACAATGAACAGGAGTCTACCAAGCGAGAGGCTAGCGGGTGGCCCTCCAAGA 900
TAAAACAACAGCATGTACAAAGGCATGGAGACATACACATCTTGACTTTTCCAGGAATGGTGGGAACGCTGGTGGAGCTAGAATGTAGGTACATAGCATA 1000
AAGTGGCAGACGGGAAGCCTTTGGAAATCTTATTACATAGGACCCTGGATGCCATTCCAATGACTTTGAATTTTCTGTAGGCTGCCAGCGAAATTTCCAA 1100
GCGTGATAGAGTCATGTCTATCTATGCACTTCAGAAAGACAACCTCAGGGTTAATGAAGAAAATGCATTGGAATATAAGAAACTGGTGACCAGAGTGATC 1200
AATTGCATGACTGTTGTGAAAGTCCAGGTGAGGGGAGCTGTGGGCAAGGTCAGAGTTGAGAGGCATTTCAGAGATAAAATGACAGTAACTAAGTAGATGT 1300
CAGACTGAGAAGAAAGGGCTGTACCAGATATATGGTGCTATCATTAAGTGAGCTCAACATTGCAGAAAAGGGGTAGGTTTGGTGGGAGTTGCTCACAAAA 1400
CATGTTTAGTCTAAGCAAAACCATTGCCATGGGCTCAGATAAAAGTTAAGAAGTGGAAACCATTCCTACATTCCTATAGGAGCTGCTATCTGGAAGGCCT 1500
AGTATACACGTGGCTTTTCAGCTGTGATTTTGTTTGATTTTAGGGATTATTCTTTTTCTGAATCTGAGCAATGTTAGCGTGTAAAATACTCACGCCCACA 1600
GCTTTGACTGGGTGAGAAGTTATCATAAATCATATTGAGTTTGTTGTGATACCTTCAGCTTCAACAAGTGATGAGTCAGGTCAACTCCATGTGAAAGTTC 1700
CTTGCTAAGCATGCAGATATTCTGAAAGGTTTCCTGGTACACTGGCTCATGGCACAGATAGGAGAAATTGAGGAAGGTAAGTCTTTGACCCCACCTG**ATA** 1800
ACACCTAGTTTGAGTCAACCTGGTTAAGTACAAATATGAGAAGGCTTCTCATTCAGGTCCATGCTTGCCTACTCCTCTGTCCACTGCTTTCGTGAAGACA 1900
AGATGAAGTTCACAGTGAGTAGATTTTTCCTTTTGAATTTACCACCAAATGATTGGAGACTGTCAATATTCTGAGATTTAGGAGGTCTGCTTCTTATGGC 2000
CCCATCATGGAAAATTTGTTTTAAAAAAATTCTCTCTTCAAACACATGGACACAGAGAGGGGAACAACACACACTAGGTCCTGTTGGGGGGTGGAGAGTG 2100
AGGGGAGGGAACTTAGAGGACAGGTCAATAGGGGCAGCAAACCACCATGGCACACATATACCTATGTAACAAACCTGCACGTTCTGCACATGTATCCCTT 2200
TTTTTTAGAAGAAGAAATAATGAAAAAAAATCTTTTTTCTATTTATATAATCATGGCATTTATAAGCATCTCTATAGAGAAGGATAATTGTGCTGAGATT 2300
AGACAGCTGTCTGAGCACCTCACACTGACCTATTTTTAACAAAATGACTTTCCGCATCACCTGATTCCGGTCTCCATGCAGGGTAAGCAGTTCCTAAGCC 2400
CTAGAAAGTGCCGATCATCCCTCATTCTTGAATTCCTCCTTTTATTTACCAAAATTCCTGAGCATGTTCAGGAAAGATGAAAAGCTTATTATCAAAATAA 2500
GTGGCTGAGATAGACTTCTTGTCACATTTGTTACAGTAAAATGGGTCTCCAAGAAAGAAAGATTTGCCTTGGGCTCTAGCATGGCCATTTATTTAAGAAA 2600
GCATCTGAAACATGAAGCTACCACAGCATCTCTCCTGTGGTTCCAGACAGAAGCCTGAGAGTCTAGGAGGAGGTGGACCGAGAAACCCTGCCAAAGTAAC 2700
TAGTAGTGCCGGGTTTCTCACAACACGATGCAAAGGGGCTAGAATCAGATGACTATTTTCATGTTTCAACATACTACACACTGGAAAACGTTACGGCAGA 2800
CTCTACTTTATAATGGGGCTGCAAATGTAAAATGACTACCTAGAACTAGGTCCTCTTAATAGCAGCAAAGTTTAAAAGGGTCAGAGGGAGCTCCAGACAC 2900
AGGTTAGATTTGATTTCTCTCCTAGTTCTGCTGTGAACAAGAGGTATAAGTTTGGCCAACTCACTTAACCCCTGAAGCTCAGTTACCTTATCTGTAAAAT 3000
GATTGCATTGTACTAGGTGTTCTCTAAAATTTCTTCTACCTCTGACTTTTTAGGAGACTAATTTTTAACTCCTTTTTAAGCTATTGGGAGAAAAATTTAA 3100
TTTTTTTTCAAAAGTTACCTTGAATCTCTAGAGCAGTTCTCAAAACTATTTTGTCCCAGGCAAAGGAAATGAGACTAGGTACCCAGAATGAGGCACCCTG 3200
CATAAAGCTCTGTGCTCTGAAAACCAATGTCAGGGACCCTGTGATAAATAATTAAACCAAGTATCCTGGGACACTGCTAGTGACATCGCCTCTGCTGATC 3300
ACTCTTGCCAGCGAGACACTCTATACTTGCTTTCTCATCATTGGCATCCAAACTGCCTACTAATCCATTGCTTTGGAAAGTTTTTTTAATAAAAAGATT 3400
ATTTCTATTAGGAAGAAAACATCCCATGTTAAATAGGAAAATTAACTGAAATCATTTTCAGATGTGATTTTTAGCACTTATAGCCATCTCAAACCATAGT 3500
ATTCATTTATACTATGCTATTTATTGTAAAACTTCTTTTTTTTTCCAAGGAAAATAAGATAGTTTGCTTTATTTTAAAACAGTAACTTTCTTATATTGGG 3600
GCACTGACCAAAATTCAATACTGGTACAAATATGTTACCTAGGAGGTCAAAATATGTGCCAGGTGAATTTTCTGAATTTCTCTAAAGAGAGAATTTTAAA 3700
CCTTATAAAACAATTAGAAACAAGTGAGTGAGAGGTGAGCATCAACAACCTGTGTAACATAAGCCACAGTACAAATTTAAGCTGAATAACCAAGCCATGT 3800
CAGTTATCCCAAATCATTTTTGTTAATATTTAGGAGGATACACATATTTTCAATAACTTAAAAGTGAATCTTTACTCCTATCTCTTAATACTCGAAGAAG 3900
TATAACTTTCTTCTTTTACTAGATTTAAATAATCCAAATATCTACTCAAGGTAGGATGCTGTCATTAACTATAGCTGAGTTTATCCAAAATAGAAAAATC 4000
ATGAAGATTTATAAAGCATTTTAAAAATAATCATTTATAGCAAGTCCTTGAAAGCTCTAAATAAGAAAGGCAGTTCTCTACTTTCTAATAACACCTATGG 4100
TTTATATTACATAATATAATTCAACAAAACAGCATTCTGACCAATGATAATTTATAGGAAATTCATTTGCCAAGTATATGTTTTATTATAAAGTTAATAT 4200
TTTGACCAATCTTAAAAATTTTTAAACTCTATTCTGACATTTCCAGAAGTATTATCTTAGCAAAGTCATCTTTATGATACCACTTATTAAACTGAAGAGA 4300
AACAAGATGGTACATTCTGGGTTTTACTTTAAAAGGGATTTGATTCAATAATTTGATTTATCACTACTTGAAAATTACATTTTCTTCCTCAGTACTGGAT 4400
GGCAATGAGATGAAAGCAGCTTTCCTGGCTCTCAACTTCCCTTCTTCATCAATTTTTCCAGCGTTTCATAAGGCCTACACTAAAAATTCTAAAACTATAT 4500
ATCACATTAATATAATTACTTATAATTAATCAGCAATTTCACATTATCGTTAAAACCTTTATGGTTAAAAAATGCAAGGTAAGAGAAGAAAAAAACACAT 4600
TGAACTAGAACTGAACACATTGGTAAAATTAGTGAATACTTTTCATAAGCTTGGATAGAGGAAGAAAGAAGACATCATTTTGCCATGTAACAGGAGACCA 4700
ATGTTATTTGTGATTTCAG**ATTGTCTTTGCTGGACTTCTTGGAGTCTTTCTAGCTCCTGCCCTAGCTAACTAT**GTAAGTCTCACCTTTTCAAGTTTGCTA 4800
CCAAAATGCATTTGCAAGGAAATGTGATATTAAATCACTCTCAATCTCTTATAAACTTCAG**AATATCAACGTCAATGATGACAACAACAATGCTGGAAGT** 4900
GGGCAGCAGTCAGTGAGTGTCAACAATGAACACAATGTGGCCAATGTTGACAATAACAACGGATGGGACTCCTGGAATTCCATCTGGGATTATGGAAATG 5000
TAGGTAGTCAACGTGCAATTTTCACTTTATTGTTTAAAAATACGATTTCTTTTTAACAAAAAATGTGCATGTTAACCATAAAGAAATTAAAAATAAATTC 5100
TAATTACACATAGCATACAGTTATAAGTAAAGGTGACCATTTTGCTCATCCGATTTTGTTCCCTAGAGATAACTACTGTTAATAAGTGTTGCATGATCAG 5200
TTAAAATTCAAACCAACAAACACTATGTTCAAGGGATTGTGGGTATATACAACAAATATGAACATCCTTTTGCCTTGCCTGCAGATACCCTCAATAATGC 5300
TGAAAGACTTATACAACATTACTGCTTCCAAAGCTTAGACTATCTCACTTTGTTTTCAAAGGAGGTTTTACGACCTTCTAAAGAGATTGAAATTGACATT 5400
TCACCTAAAACTCGGGGAAATGTAAATGACAATATTAATTGGTAAGAGAGGAAAGAAGAAAGAAAGAAGGAAGGAAAGAAAGAAAGAAGGAAGGAAGGAA 5500
AGAAAGAAAGAAAGAAAGAAAGAGAGAGAAAGAAAGAAAAAGAAAAAAGAGAGAAAGAGAGAAGGAAAGAAAGAGAGAAGGAAAGGAAAAGAGAAGCAAA 5600
GAAAGAGAGGAGCAAAGAAAGGAACACTTAGCACTAGTTAGGAGACCCAACTCTGGAATTATCAGCTATATATTTAACAAACGTTATACTTTTAAATAGC 5700
AAACTCTTTATTGTTTCAATTTTATCTGGTCAATTGGAAAAATAATTTTTGTCTTATCTGTCTCCTTGAAATGTGAGGATCAAAGGAGACTAAAACATGA 5800
TAGCTTTTAAAGTCTATTTCAGTAAAACAGACTTATATAGAGGGGTTTTTATCATGCTGGAACCTGGAAATAAAGCAAACCAGTTAGATGCTCAGTCTCT 5900
GCCCTCACAGAATTGCAGTCTGTCCTCACAAATGTCAGCAATAGATATGATTGCCAAGCAGTGCCCCATCCAGTGCTCTTATCCCAGCTCATCACGATCT 6000
TGGAGTTCCCATTTCTCTCTGCAGGTGGAACTGACCTCTGATAAGAAAAGCTCCTCGGAGAACACATGCCTCACTATTTGCCATCTACTTTAACAG**GGCT** 6100
TTGCTGCAACCAGACTCTTTCAAAAGAAGACATGCATTGTGCACAAAATGAACAAGGAAGTCATGCCCTCCATTCAATCCCTTGATGCACTGGTCAAGGA 6200
AAAGAAGGTAAAAATAAAGGCTTTTTATTTTTGGTGAGGGGAGAGGTTTTACATCCTTCAGTAAATAACGAGAAGATCACAGTCATTCCCTCTTGACTA 6300
CAGTATGTTGTAGTGTGCAGCACAAAGGGGGAAGTTATTGGTGATTGCCTGAGGGAAGGCAACTTCTGCCACATCAAATGCTGTGGCTCACACCTACCTC 6400
TACAACCGCTGAGCAAAGCACTTGAAACCTTGACTGTTAGAGGAGCAAAGCTCTGGTCACACCAATAGGAGCCTCAGTACTTTGCCAAGGACATTTTTCT 6500
GCAAGAGTTAGTTAGGGTTATTAGATTTAGCAAATGAAAATAGAAGATATCCAGTTAGGTTTGAATTTTAGGTAAGCAGCAGGTCTTTTTAGTATAATAT 6600
ATCCTATGCAATATTTGGGATATACTAAAAAAAGATCCATTGTTTATCTGAAATTCAAATGTAACTGGGTATTGTATATTTTGTCTGGCCATACTAATCC 6700
AGGTGAGTGGAAAGAAGAGATCCATAATGTTTTAAAATATTTGCCTGAGTTCATATTCCTATAACTGATAAATGAGTACCTTTCATTGACAAGGTAGAGA 6800
AAATAAATAAACTGCATTCTCAGAAGATGATTATTACATAGTCTAATCCAAGGAATCTATGATGACCAAATGAGGTCCAAGTTGCAGAATAAATTAAGCC 6900
TCAGACTTCTGTGTTTATGAGAAGCTGAGGTTTCAAACCAGCTAAATCCCTTAGGACACTTAGAAATGCTAAGATATACAGAATAAGCTAGAAATGGCTC 7000
TTCTTCATCTTGATTATGGAAAAATTTAGCTGAGCAACACTCACTGTTGGCCTCGTATACCCCTCAAGTCAACAAACCACTGGGCTTGGCATTCATTCTC 7100
TCCCATTCTTCCTTTCTACCTCTCTTTTCCACACTCAG**CTTCAGGGTAAGGGACCAGGAGGACCACCTCCCAAGGGCCTGATGTACTCAGTCAACCCAAA** 7200
CAAAGTCGATGACCTGAGCAAGTTCGGAAAAAACATTGCAAACATGTGTCGTGGGATTCCAACATACATGGCTGAGGAGATGCAAGGTGAGTAGCATCCC 7300
TACTGTGCACCCCAAGTTAGTGCTGGTGGGATTGTCAGACTATCCTCGCGCGTGTCCATAGTGGGCACCAGTGATGCAGGGATGGTCATCAAGGCCAACA 7400
TTTGTGCAGTGCTTGCTCTGTGCCAGGTACTGTTCTATGTGCTTTAAGTGTGTTAACTCGGTTCTTCACAGCAATCTTATAGGTTCTATTTTAATCCTAC 7500
TTTATGGATGAGGAAACTGAGGTACAGAGAGGTCACAAAATCCTTGCCTGGGTCAATTCCAAGCATTTTGGCTGTGGATTCTGTGCTCTTAAATATTATG 7600
GAACACTGCCTTTTAAGTGTGAATCAAGAGTAGACTCAAGTCATATTCAAAAGAATGCATGAATGGCTAAATGAAAGAAGAATGCTAATAGAATCTATTA 7700
ACTTTCTATAGCTCAGACAATCACTTAATTTCTGGACATTCAAAGAACAGCTGCACACAAACAAAGTGTCTACCTAGGGACCTAACTTAATGGCAATTTT 7800
CCAGATCTCTGAATTGATTGATTTCATCACAACAAGTAGATAAACCTTGACATTAGCACATAGCTAGTTTGGAAACCCCTACTCCCCCAATCCCCTCCAA 7900
GAAAAGAGTCCTTAAATAGACATTAATATAGGCTTCTTCTTTTCTCTTTATTAG**AGGCAAGCCTGTTTTTTTACTCAGGAACGTGCTACACGACCAGTGT** 8000
ACTATGGATTGTGGACATTTCCTTCTGTGGAGACACGGTGGAGAACTAAACAATTTTTTAAAGCCACTATGGATTTAGTCATCTGAATATGCTGTGCAGA 8100
AAAAATATGGGCTCCAGTGGTTTTTACCATGTCATTCTGAAATTTTTCTCTACTAGTTATGTTTGATTTCTTTAAGTTTCAATAAAATCATTTAGCATTG 8200
AATTCAGTGTATACTCACATTTCTTACAATTTCTTATGACTTGGAATGCACAGGATCAAAAATGCAATGTGGTGGTGGCAAGTTGTTGAAGTGCATTAGA 8300
CTCAACTGCTAGCCTATATTCAAGACCTGTCTCCTGTAAAGAACCCCTTCAGGTGCTTCAGACACCACTAACCACAACCCTGGGAATGGTTCCAATACTC 8400
TCCAACTGCTCTGAAGGCTGCTCTATAGTTGTTGTGCACAGTAACGTTTCAGATCAGCAAGGACAGAAATTCACTCAGAGTCACATAAGCAAGAAAGGGT 8500
GGTTTATTGCAAGGATATAAGAATATCTCAAAAAATCTGAGGGCAAGAAAAGCAGCTGGTTATCAAAAAAACTACATCCAGGAACTCTGCTGACACCAAG 8600
ACCCAAGTCTGCTTCTCTCCCTGAAGCCACGCGTTACCTTTTCTCTCTTCTTCTCTAAATTTTACTTCCCTCTTTTTCGCATATTGCGGATGGATTTTTT 8700
TGTGCTCATGGCACAGAATGGCTGACCCAAATTTTCAGCCTCATCTCTCGTGTTTACATGTCTTTCCAGTTCAAGAGGTAATAACTCATATTGAATCCCA 8800
ATTCCATTTTCCTTGATCCTGTTTGGCTCAAAGTTAGCAACTTCCATTCACTGATGGCCAGGAAAGTGAAAGGTCTCAGAGAAAATAGAGACTCTAAATA 8900
TTGTGTATTACAATATAAAACTGTTAACTTTTCAAAGCTGCCCCCATACATCTTTTTTTAATCCCCAGTGCCATAG 8976

Figure 5.11: Strategy view of SeqMan assembly of the mouse genomic sequences

This is a “strategy view” diagram of the final assembly of the full-length mouse genomic sequence. This was achieved using Ensembl’s Trace Archive of Whole Genome Shotgun sequencing reads generated by the Mouse Sequencing Consortium on-line at http://www.ensembl.org/. This database was interrogated with the mouse foveolin cDNA sequence. Numerous matching mouse genomic sequences emerged and were down-loaded and assembled using SeqMan. Sequence from the overhanging intronic ends of these assemblies was used to re-interrogate the databases. Eventually the full-length mouse gene was obtained, with the final assembly containing 68 separate sequences, of which the longest represents an earlier, slightly shorter, assembled consensus sequence.

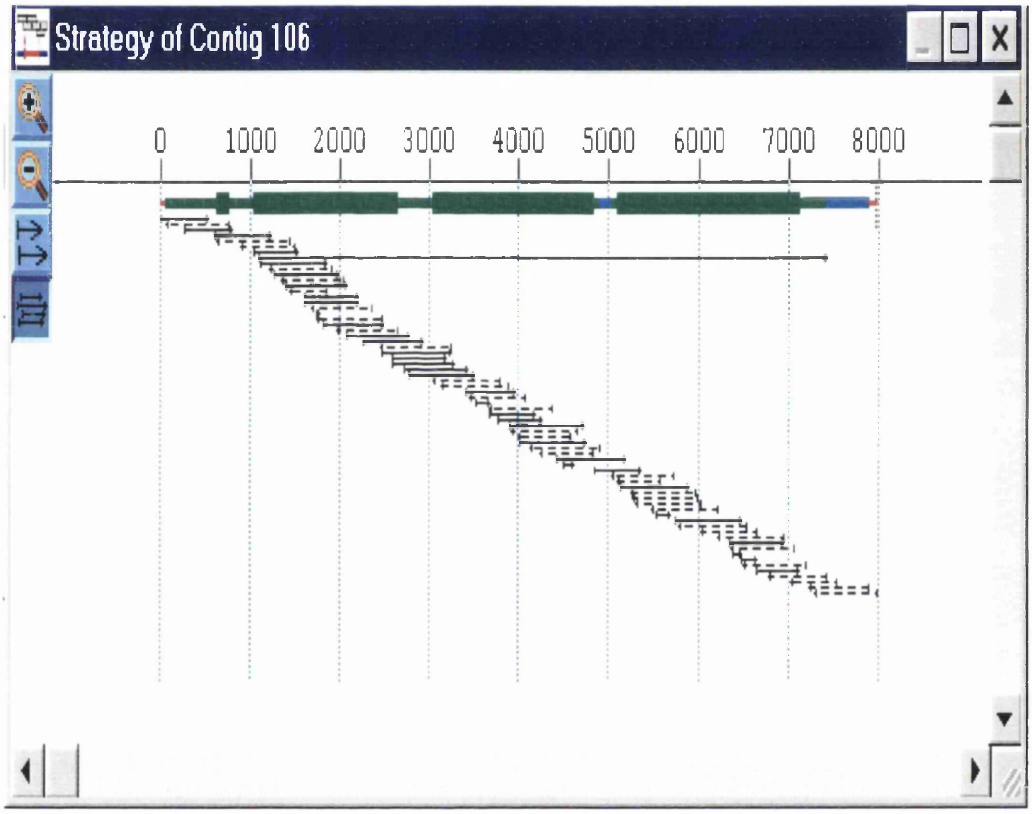

Strategy of Contig 106
0
1000
2000
3000
4000
5000
6000
7000
8000

Figure 5.12: The full sequence of the mouse foveolin gene

This figure shows the mouse genomic sequence of foveolin. This was obtained by interrogating the Ensembl Trace Server mouse genomic sequence databases and gradual assembly of the resulting matching sequences, as described in the legend for Figure 5.11. The exons were identified by comparison with the known mRNA sequences and are highlighted in yellow. It is possible that the first exon starts slightly more upstream, as discussed in the legend to Figure 5.7.

ATCTAGAGCAGTGGTTCTCATCCTGTGGCCATGAGCCCTTTGGGGGGGTTGAACGACCCTTTCACAGGGGTCACATATCAGATATCCTGCATCTTAGCT 100
ATTTACATTATGATTCATAACAGTAGCAAAATTAGTTAGGAAGTAGGAACAAAATAACGTTATGGTTGTGGTCACCACTATGTTAGAGGGTCGCAGCATT 200
CAGGAGGGTTGAGAACTGTTGTTCTAGAGGCAAATAAGAAGACAGAGTTCCTTGATAGGGCCCAGAGGCAGTGAAAGAAGTTTCCACGTAGAAAGTGAAG 300
AAGGTCTGGTGTCCGAAGCAGTGAGGAACTTAAAAAAAGAAAACCAAAAACATTGCCAACTAACAGTCCAGGAGAAGAGCGGGGCATGAAAGGCTGAGTT 400
CCCATGGGATGCCTTGAATGGAATCAGAGTGTGGGAAAATTGGTGTGGCTGGAAGGCAGGTGCCGGGCATCTCAGACGCTGGTAGCTGGGGAAACAGGAA 500
ACCCCTTTAGGATCCCAAGATGCCATTCCAATGAGCTTGAGATTTTTCTCATGGACTGCCAGTGAATGTTTCTACGCTCCGGAAATTAATGTTTACTKAT 600
KTTCCATATTCTAGGGGAGAACCSTGGGARAAATGGAGGACATTCATTGAAATATCTGAGTCCTGGGATAAGGCAGGCTTGGTCCTACAACTCTGGTAAA 700
AGTCCATCAGGAAGTGCCTTGACCAAGGCTGGAGTGGAGAGCTGTTGGTGAGATGTAAGGGCAAGGTTTAGTTGCTAGATATGTAGATGGCAAGATGGTG 800
CTGCCAACAGCCCCCAGAGCTCTAACCCACTGAGAAACCCAGGAATGAATGATGGGAGATGGCTTTGGTGCCAGCTGCTAGTGACATGGCTGGAAAGCTG 900
CACTGGCTTCGAGGCCAGACAATTCCTCAAGGAAACATCTGGCCAGGGTGCAAGGGCCAGTTTCCTTCCTTGGAGTTCCTTTCACAGCTAAGAACATCAT 1000
CCCCCAACCACTGGTTTTGTTAAAAAGTTTTCAGTATGACTTGAGCATGGTCAAGAAGCATAGAGAGGGGGAAATAAGGGTGGAAGGAGCTGGAGAAAGC 1100
TTACAATAGGACTGGGTAAAGGGAAGGAGAAGAAACCATTCCCGCATTCCCATAGGAGCCAGTACCAGGAAGGGCAGGTGTACACACAGATCTCATCTAA 1200
GGCCATGTTTGGTTTAGGGATTACTCTTCTCCTGAATCTGAGCAGCAGCAATACGTAAAATACCCACACCCATGGCTTCCATATTCCAGAACTTATCACA 1300
AACCGTGTAGAGTTTACTGAGATACCTTCGTCAGAGGATGAGTCAGAGGCCTCCTGCCTAAGGGCCCTACTGAGCAGGCAGCTAAAGGCTTCCGGGCCTC 1400
TGCAGCTCCACAGATACAGGAGAGGGAAGCAGATAAGCCGTGGACTCCACCTGAGCACACCTAGCTTGAGCAAAGCTGGTCAGGTACAAATAGCAGAGGG 1500
CTGAATGTCTGT**GAGCACGCCGCCTGATCCTCTGCTCCACCACACTCCTGCCGCCATGAAGCTCACA**GTAAGTCAGATCTTCCTTTTCAATGCAGCACCA 1600
TACAACATTAATAGTCAGGGGTGAGGGGGTCTGACTCTTACGGCACTGTTACCATAGTGGAAATATTCTCCTTCCTTTTCATGGAATCATGGTGTTTACA 1700
AGCATGTCCATAGAGAAGAAGAATTGCCCCGGAAGAGCCTGTCACAGGCTGAATACTGTAGAATTGTCTTTCACACCATCTGTTCCAAGGTTCTACTTAA 1800
GACGAGCAGTCTCTGGGCTCCAGAAAGAGTCTTTCTTAGCCTTGATCTCTTTCTTATTTCTGATTTCTCCTTTCTTATCCATGATTTCCACTTTTACCAG 1900
TTCTGGGCATGTTCCGGTCAGACTGGAAGATCACTGTTGTCAAAACTAGTCTTCAACACTCTTGGCTGTTAACATGAAAACAACGGTCCTTGGGCCCTGT 2000
GCAAGCATTTCTTGGAGAAAGTCTCTGGGGATGAAGCTATCTCAGTTTCCCCACTGAAGTCCTAGGATACAGAGGCTCAAACAGAGTGCACATATTCAAT 2100
TTCAGCATACTCTATTGGCGCTGCTTTATGAATCATATGAATTTATGGAATTGGAAATGTAAACTATGACCAAGAAGCGTCCACCTCAGAACAGGTTGGG 2200
TGGGGAACTCCAAGCACAGGCCAGAGGGCTGCGTTTCTCTTCTAGTTCTGTCTAGAGGAGTGATTCTCGACCTTCCTAATGCTGTGACCCTTTAATACAG 2300
TTCCTCACGTTGTCGTGACTCCCAGCCATAAAATTACTTTCATTGCTACTGCATAACTGTAATTTTGCTACCATTATGAGTTGTAATGTAAATATCTGAT 2400
ATGCAAGATACCAGATAACCTAAGAAACGGTTGTTTGACCTTTAAAGGGGTCACAACCCACAGGTGGAGAACTACTGGTCTAGGGTCCTTTACAGTCCTT 2500
TAGCTGCCTCATTTACAGGAGATAACATCATGCTCAAAAACTCCCTCCACATTTGGCTTTTTGGGTTGTTTTGTTTTGTTTTTCAAGACAGGGTTTCTCT 2600
GTGTAGCCCTGGCTGTCCTGGAACTCACTTTGTAGACCAGGCTGGCCTCGAACTCAGAAATCCGCCTGCTTCTGCCTCCTGAGCGCTGGGATTAAAGGCG 2700
TGCGCCACCATGTCTGGCTCACATCTGGCTTTTTAAGAGACCGATTTTAACTCTCTTGCATTGAAAATAAATATAGTAGAAATGCTTAACCTACTAAGAC 2800
AATAAAAACAGGATTCCTTCTGCTAGGAAGAACACGTTCCAGACTAAGGAAAAAAACCTTTTCAGGGCTTTCATTACACTGTGCCATGCACTAATTTTAT 2900
GTTTTCTTCATCAGTTTTCAGTGTCTGAAATTCAGTGTCAAAATTCTAAGACTACATATGAATATCATTACAGTAACTCAGCAATTCTATGTTACCAGTA 3000
AGTTTTTCTGTAGTTTAAAAAAAAGGTGGAAGAAGAAAGCACAGATGGTTTAGCACATGGGTAAAATCAGTAACTATTTCTGATGAGCTTGGTGAAGATG 3100
CTGTAAACCATGCGACCACCAGTCCTGTTCTCTGTGCTTTCAG**ATGTTCGTCGTGGGTCTGCTTGGCCTCCTTGCAGCTCCTGGTTTTGCTTAC**GTAAGT 3200
CTCATTTTTCTGAAGTTCATTGTCAAAACTGCATTTACAGTGAAATGTGATCTTAAGTCACCCTCTGCTTCTTATGAACATTAG**ACGGTCAACATCAATG** 3300
GTAATGATGGCAATGTAGACGGAAGTGGACAGCAGTCGGTGAGCATCAATGGTGTGCACAACGTGGCCAATATCGACAACAATAACGGCTGGGACTCCTG 3400
GAATAGCCTCTGGGACTATGAAAACGTATGTAATGGACACACAGGGTAAAGATATGGTGTAGCCACCACCCATTAAAATTTCTGAGGTGAATTCTAGCTG 3500
TTCATGAACATTAAAAGCTACTAGTAAAAGTGCCCATTCCACTCAAAACAATTTTACTTTTTTGCATATAATTATTGCTAATAAGTATTACACAATAGGT 3600
CGAAATTCAAAGGGATCAATAGTAAGGATAAAAACTATGTACAAAGACAAACACAGCATCCTTTGGTCTTCCCTGCAGAGAGTCTCCATGATGTTAAAGG 3700
TCCAATGTTTTATGGAGGCTGAATGAAATACGAATGCCTCTGTGATGGAAAAGGCCCAACATCTTATGGAGAATGAGTGAAGTATGAATGCTATTAGTTG 3800
TAAGAGAAGGCGATGCAAAGCAACACTTGGCACCACCTGCCAATTACTACTTTCCTATTTAAATGTAGTTTAAAAAGCAAAGCCTGTCTTCCCTGCCTCC 3900
TGGAAACACTGCGGATGGAGGTAGACCAAGGTATGACAGCCTTTAAAAGTTTGTCAGCAAAACACTTCCCCCATACACACATACACACACCCTCCTACTA 4000
CACTGGAACTGAAGCAAAGGCAGTGGGTTAGATATATCCACCCTCTAAGAGTTTGCAGGTCATCTATATATGATAGCCAGAGACACAACTGCAGGACAGC 4100
CAGACTCTGAGCACTCTCCCCAGCTCCTTGTAGCTCTGTTTCAGTGGTGACTTGTGACAAGAATCCTGGGGAACCTGTGCCTCACTGTTCTCTGTCTTCT 4200
TTAATAG**AGTTTCGCTGCCACGAGACTCTTCTCCAAGAAGTCATGCATTGTGCACAGAATGAACAAGGATGCCATGCCCTCCCTTCAGGACCTCGATACA** 4300
ATGGTCAAGGAACAGAAGGTAAAGTCCTGCCTTCTTCTTTGGAGTGACAGGAAGTCTTACAGTCTCCAGTACACAGTGAAGTCACCCCCATTCCCTCTTT 4400
GGTGGAGCATGACAGCATGTTTGTCATGATAAATGCCACAAACATGTAAAACTGTTCAGTGTCTGCCTGAATGGAGGGTGGCTTCCACTGTGTCAGATGC 4500
CGTGGCCCACATCTGCCTMTGCAGGGTTCCAGTAAAGCCACTGGCTATCTTGAGTGTCAGAGTACCCAAAGGTCTGCACACTTCAGTACAAGCCCTCCAT 4600
ATTTCAAGGGCACACTCCTACAGTCGTTGGGGTTATCAGAACTAGCAAACATAGAGACTGGATTTTCAGATGAAAAGAAATCCTTTTTAAAGTCTAAGTA 4700
TGCCTTATACAATGTTTGAGATATTCTCAATACTAAAAAAAAAAAATTGTTGCTTGCTTGAAAATCCAAATGTAACCAAGTGTCCTATATCCAGTGTCAA 4800
TCATGGCTGTAGTAGATGGGAAGAGGGAGCCCGTGGTTTTCACAGTCAGACGCCTGAGTTATTCTTCTAAGTGATAAATTGGTTCCTATAACAAGCAAGC 4900
TAGTGAATATAAATAAGCTCTATCTCAGAAGTTATCCTGTAGTGCTACCCTAGAATCTAAGAGAGCAAAAGTGCTTCAAATTTCAGAATAAGTTTTGCTT 5000
TGGACTTCTGTTTTTCTAAACAACTATAACTTCAAACCATCTAAGCCTCGTGGGACACTTAGAAATACCAAGCCATTCAAAGCTAGAATTGTTTCTTCAC 5100
CTTACTTGAAAACAAAATGACAACCAAAAATTGTCCCCACTGCCCTTGTACATCTTCAGATCAGTAAAGTCCTGGGCTCAGGGATCATTCACTTTCTTTC 5200
TTTCCTTTCACACTCAACTTCAG**GGTAAAGGGCCTGGAGGAGCTCCTCCCAAGGACTTGATGTACTCCGTCAACCCTACCAGAGTGGAGGACCTGAATAC** 5300
ATTCGGACCAAAGATTGCTGGCATGTGCAGGGGCATCCCTACCTATGTGGCCGAGGAGATTCCAGGTGTGTACCCTGAGATGCTGTATATCCCAATGCAG 5400
TACTGAGAGAGCCATCAGACACTCTAAAGTGTGACCACAGACGGACCAATCATGTGGATTATCAGAGCAAACACTTGCTTGCTCCTTGTCAGACAGTTGT 5500
CCATGCTTCAAAAGTTCATTAAAAAAAATAGTTCACAGGCTCCTCACAGAAACCTTAGTAGAATCCACAGCTTCTGCTCTTAGTCTTACTTTTAGAAAC 5600
TGAGACCCAGAGAAAGGTCACAAAACTTTTGTCTGGCTCAGGTTCTATGTCTTTAACTTTATAGAATACCGTCTTTCTGGGTGGGTGGGCTCTAGAGTAA 5700
ACTTCAAGTGAGTTCAAGGAAAGCATGAGAAGTAGGGAAGACCAAATGAAAGGAGAATGCCAATGAAATCTATCGATTCTATAGCGCCAATGCTTAACTC 5800
CTAGGCGTTCAAAGAATAGTATCCACAAGGTGTCAGCCTAAGATCCTAATCTAACAGCAAGTTTTCAGATCTCTGAAGTGAAAAGAGAAAGCAAGAGAGG 5900
AACAGAGACAGAAACAGTAAGAGACAGAGAGGCAGAGACAAAGAGACAGGGAGAATAGAGAGGGATTAAAATTAATATATAGTTTAGAAATTACGACTCC 6000
TCACAGTCCCTGCAGAGTCCTAGGATAGGCACTGATTTGGACTTCTTTTCTTCTCACTAG**GACCAAACCAGCCTTTGTACTCAAAGAAGTGCTACACAGC** 6100
TGACATACTCTGGATTCTGCGGATGTCCTTCTGTGGAACATCAGTGGAGACATACTAGAAGTCACAGGAAAACAACCCGTGGGCTCTGACCATCGCAATG 6200
CTTGATTATGAGAGTGTTCTCTGGGGGTTGTGATTAGCTTCTTTAAGGCTCAATAAACCCACGTGGCAGCACATCCAGTTTGTAATGACATGCCTCATGA 6300
CTTCTATGGGAGTCCAATGTGGCACCTGCCAGCCTGTATTCAGGACCTCTCCGCTATAAAGCATCCCTCCAGAGTTTTCAAATACTACAAAGCACAGCCT 6400
GGGTTTGGGCTCAGATAGGCCACTGCTGCCTGACTACATTACAGACAAACAAGTTTTAAAAGAAAGAAAAAAGAGCTCAGAGTGGCTGGAATCAGCAAGG 6500
GTGTTTTTCCTGCAAGGAGCCAGAAGTATCAATAATCACCCAAGGAGGAGACACTGGGAATGAGAGACTAGAACACACGCCTGCAGATACGGAGAACCTC 6600
AGATTGCCGCTCTCTCCCATAACTGCACACCCCCTTCTGTAAACTCTGCTTCTTTCTTTCACCTGAAGATGGCCCTTGCTTTTTTTTATTATAGGACAGG 6700
ATAACTAGACCAGAAAGTCAGCCTGACTCTCTACATTTATATGTCTTCCCAGTTCAAGAAATATTATTTACTGTTGAATGGCACTTCTATATTCCCTTGG 6800
TTCAATAGTCTACAGGATCCATCCATTGACAGGCCGAAGAGTGAGATCACATGATACCCAAGCACATGGGTCTTCCCTGAAGGAGAAGGATCCAGAAGGA 6900
GGTAGAGTCTTGATATTGTGCATTATAACAGGAACCCACACTTCGTTTTAAAAGTTGCTCCCATATGTTGTGTGGTCCCTAGAATCATCAGAGAAATCTT 7000
CCTCCATTCTTTTATGCTTACATGAGCTGTCTCTATGTAAGCCTAACAAAGTGAGCAAAACGTTGATTCCCACCACTTAGACAGTCATCTATATTGGTCA 7100
GTGTCTACCAGAAGGAGGTTTTCATAATTATCTTCATTATGTTGCAATTTCATCACTTTTTGCAAATTAGTGTTATTTCTGAAATATTCTGAAGCCCCTA 7200
AAGATCGCAGTGGTAGGAGGCTGGAGAGATGGCTCAGTGGTTAAGACCACCCACTTTTCCACTCTTCCAGATAACCCAGGCCCAACTGTCAGTCCCCACA 7300
TAGGAGCTCACAAACCATCCCTAACTCAGGTTCCAGGGGATCCAAAATCTTCTTCTGGCCTCTACACACACTAGACACATATGTGGTGCACTGACATTCA 7400

Figure 5.13: Comparison of the structure of the human and mouse foveolin genes

This figure compares the exons and introns of the human and mouse foveolin genes. The exons have a black background and the introns have a grey background. Both genes contain six exons, of similar length, again emphasising their high degree of homology. It is possible that the first exon of the mouse gene starts slightly more upstream, as discussed in the legend to Figure 5.7.

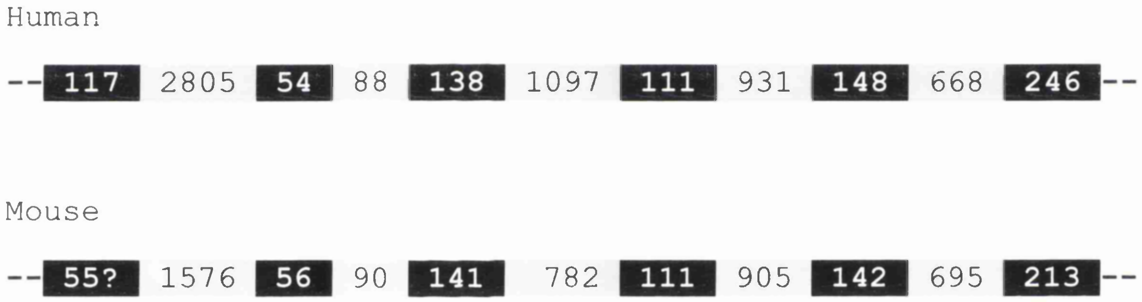
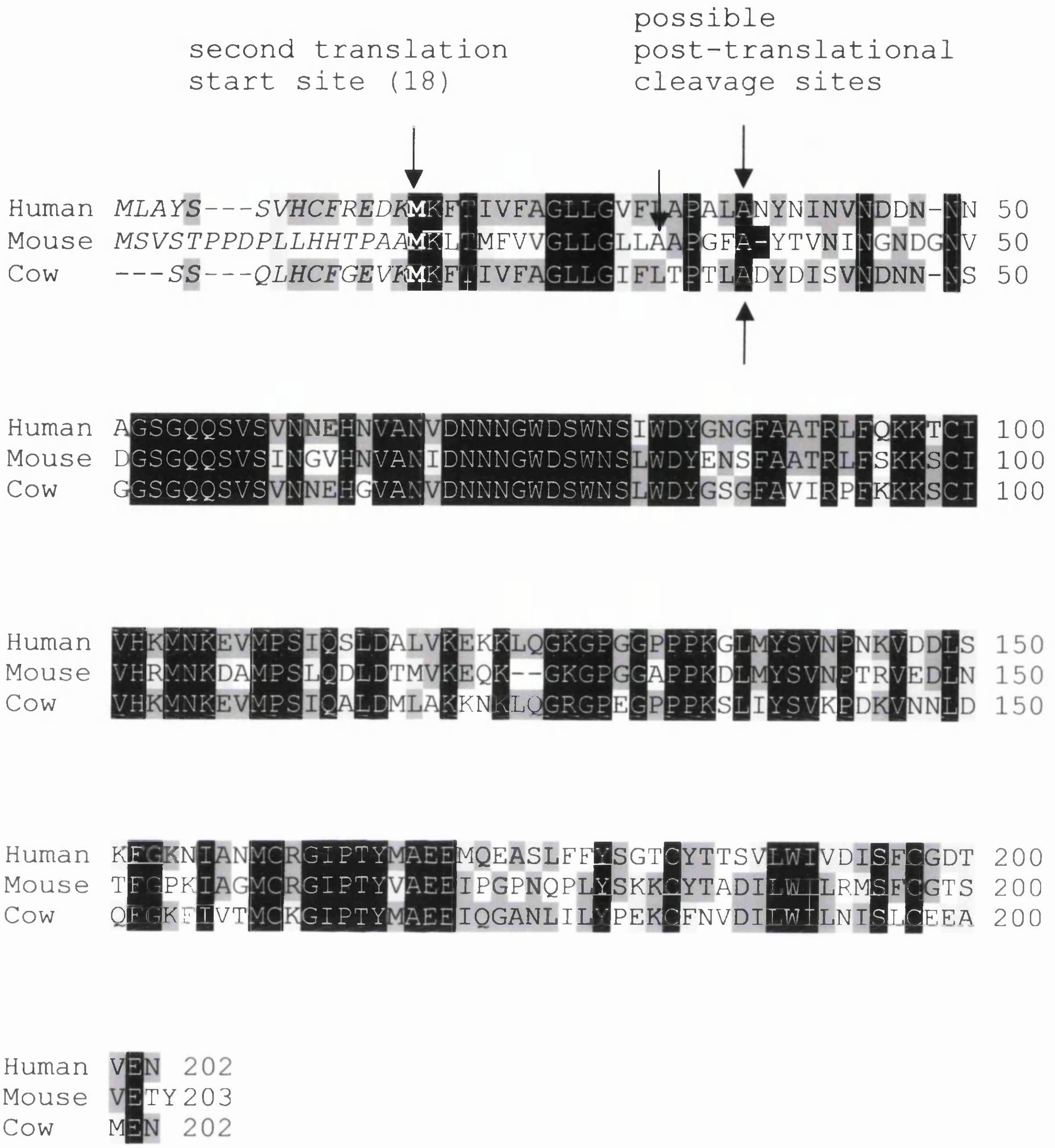

Human
--
117
2805
54
88
138
1097
111
931
148
668
246
--
Mouse
--
55?
1576
56
90
141
782
111
905
142
695
213
--

Figure 5.14: Predicted primary sequence of the human, mouse and cow foveolin proteins

The Open Reading Frames (ORFs), or translatable protein sequences, within the three human, mouse and cow mRNAs were translated using EditSeq. Here, the resulting primary protein sequences are presented and compared. As discussed in the text in Section 5.1.3, the translation start site used (amino acid position 18 with M in bold type, arrowed) is the second possible site for the human sequence. For the mouse sequence, the translation start site used is the main one in the on-line sequences but the second possible from my RACE results, which included further upstream nucleotides. For the cow sequence, the translation start site used is the first and only one found.

Amino acids which are identical in all three sequences have a black background; those which are identical in only two of the three sequences have a dark grey background; and those amino acids which are regarded as similar in terms of their physico-chemical properties, being, for example, polar or acidic, have a light grey background. The high degree of homology between the three sequences is clear, although the upstream amino acids, in italic type, are much less similar.

Using the Network Protein Sequence Analysis programs on-line at the Pôle Bio-Informatique Lyonnais, all three proteins are predicted to contain a starting signal peptide. The likely post-translational cleavage site for this signal peptide in each protein is indicated by an arrow.



```
Human  MLAYS---SVHCFREDKMKFTIVFAGLLGVFLAPALANYNINVNDDN-NN 50
Mouse  MSVSTPPDPLLHHTPAAMKLTMFVVGLLGLLAAPGFA-YTVNINGNDGNV 50
Cow    ---SS---QLHCFGEVKMKFTIVFAGLLGIFLTPTLADYDISVNDNN-NS 50

Human  AGSGQQSVSVNNEHNVANVDNNNGWDSWNSIWDYGNGFAATRLFQKKTCI 100
Mouse  DGSGQQSVSINGVHNVANIDNNNGWDSWNSLWDYENSFAATRLFSKKSCI 100
Cow    GGSGQQSVSVNNEHGVANVDNNNGWDSWNSLWDYGSGFAVIRPFKKKSCI 100

Human  VHKMNKEVMPSIQSLDALVKEKKLQGKGPGGPPPKGLMYSVNPNKVDDLS 150
Mouse  VHRMNKDAMPSLQDLDTMVKEQK--GKGPGGAPPKDLMYSVNPTRVEDLN 150
Cow    VHKMNKEVMPSIQALDMLAKKNKLQGRGPEGPPPKSLIYSVKPDKVNNLD 150

Human  KFGKNIANMCRGIPTYMAEEMQEASLFFYSGTCYTTSVLWIVDISFCGDT 200
Mouse  TFGPKIAGMCRGIPTYVAEEIPGPNQPLYSKKCYTADILWILRMSFCGTS 200
Cow    QFGKFIVTMCKGIPTYMAEEIQGANLILYPEKCFNVDILWILNISLCEEA 200

Human  VEN 202
Mouse  VETY203
Cow    MEN 202
```

Figure 5.15: Predicted secondary structure of the human, mouse and cow foveolin proteins

The secondary structure of the three foveolin proteins has been predicted as a consensus of the results from twelve methods within the Network Protein Sequence Analysis programs on-line at the Pôle Bio-Informatique Lyonnais. The thin yellow lines are random coils, the mid-sized red lines are extended strands and the thick blue lines are alpha helices. Again, the three proteins are similar but unfortunately it is not possible to extrapolate from these secondary structures to arrive at the likely functions of the protein.

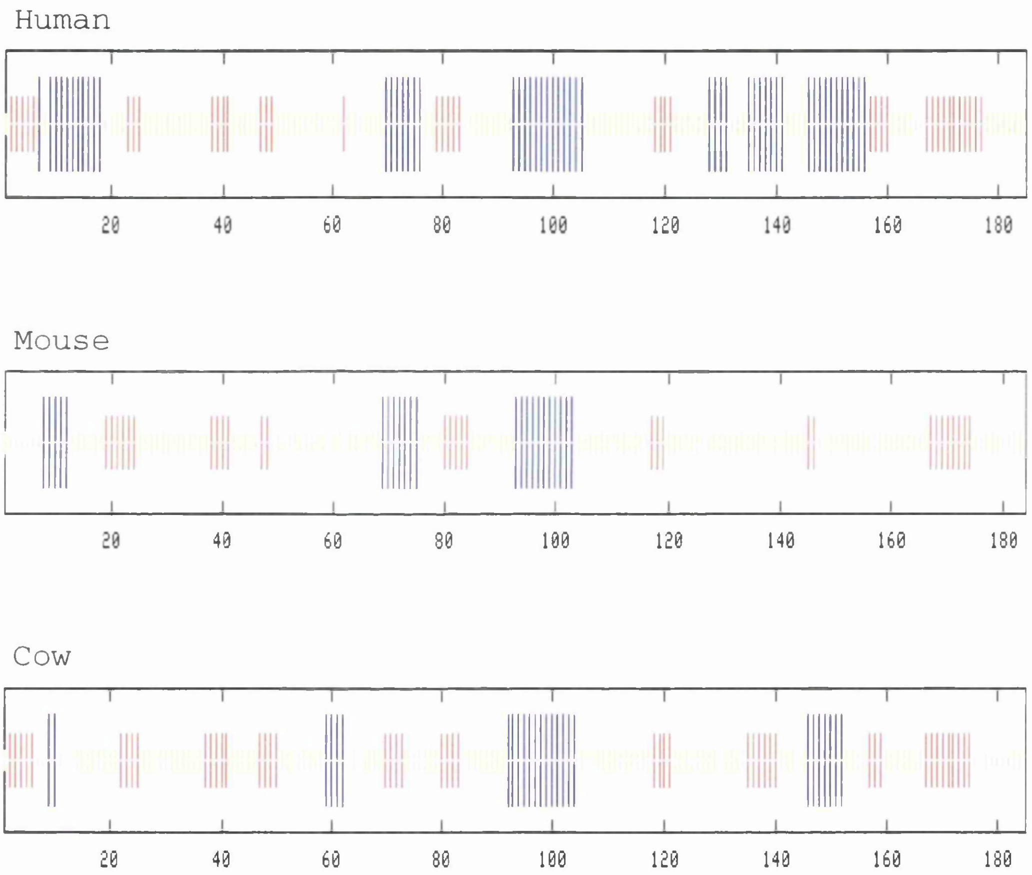

Human
20
40
60
80
100
120
140
160
180
Mouse
20
40
60
80
100
120
140
160
180
Cow
20
40
60
80
100
120
140
160
180

Table 5.1: Physico-chemical properties of predicted foveolin proteins

This table shows the physico-chemical profiles of the foveolin proteins from human, mouse and cow, predicted using the Network Protein Sequence Analysis programs on-line at the Pôle Bio-Informatique Lyonnais. The human protein is that predicted from the second ATG site. It is clear that the properties of the three proteins are very similar.

Information on predicted foveolin protein	Human	Mouse	Cow
Molecular Weight (Daltons)	20330.04	20132.97	20519.59
Amino Acids	185	184	185
Strongly Basic (+) Amino Acids (K,R)	15	16	17
Strongly Acidic (–) Amino Acids (D,E)	17	17	19
Hydrophobic Amino Acids (A,I,L,F,W,V)	61	57	63
Polar Amino Acids (N,C,Q,S,T,Y)	59	55	52
Isoelectric Point	5.726	6.341	5.773
Charge at pH 7.0	–1.875	–0.875	–1.871

6 Further overall discussion

The classical candidate gene approach to the identification of cancer-associated genes has been extremely successful. Many genes involved in the development, progression and metastasis of cancer have been identified in this way. Our understanding of tumour biology, in particular tumour suppressor genes and oncogenes, has greatly expanded over the past three decades, and the fruits of these labours are now reaching the clinic.

However, the average number of critical events affecting such genes which is predicted to lead to the acquisition of a malignant phenotype is only around six. Yet cancer cells obviously have many more changes at the DNA, RNA and protein levels than this. Overall, only a small proportion of genes changes significantly in expression in disease states, even when comparing cells at the opposing ends of the normal versus cancer spectrum. Although the exact figure obviously depends on the criteria used, only around 1% of transcripts are significantly up- or down-regulated. This nevertheless represents a large number of genes, for example around 100 in this study of gastric carcinoma and normal stomach (see Section 4.2.1).

Thus in cancers a few molecular events are causative whereas most represent their consequence. Changes in expression resulting from these cancer-causing mutations are nevertheless interesting and useful because they may still yield diagnostic, prognostic and therapeutic targets. Such multiple expression changes presumably result from only a few critical genetic events through additional changes at the DNA level and through altered gene regulation via transcription factors, methylation and other epigenomic events. Large-scale gene expression profiling through DNA microarrays and SAGE has only been widely used over the past few years but has already produced hugely exciting potential and actual targets for improved cancer screening, diagnosis, prognostication and treatment, as discussed in Section 1.7.2.2. There, the example was given of diffuse large B-cell lymphoma, in which immuohistochemistry for CD10 and bcl-6 can be used to identify patient subgroups with good and bad prognoses and therefore to direct more intensive therapy.

Clinical use of these candidates will almost certainly go hand-in-hand with more traditional techniques: for example, the expression of oestrogen receptors or c-erbB2 in breast cancer is currently assessed on histopathological sections of breast cancer tissue by immunohistochemical staining and microscopy. But the even greater numbers of potential targets now emerging need to be validated and functionally investigated both *in vitro* and *in vivo*.

As Carlos Caldas and co-authors discussed in a recent review article (Caldas *et al.* 2002), it is increasingly obvious that classical clinical and histopathological criteria used in the diagnosis and management of patients with cancer, whilst valuable, could be improved upon; and that this could be achieved by their supplementation (rather than replacement, at least in most cases) with the emerging molecular classifications. In order to improve treatments, the latter will need to be judged against tumour categories at the DNA, mRNA or protein levels. This presents a number of challenges.

Tumour material and clinical data will have to be collected carefully and efficiently. This has technical, cost and ethical implications. One technical issue is the preservation of the molecules of interest. Biological material starts to degrade as soon as it is handled during removal from the body. While degradation of biological molecules does affect DNA and protein, it is RNA which is most vulnerable, due to its intrinsically labile nature, as already discussed in Section 3.1.2.1 (Liotta *et al.* 2000). When specimens are intended for traditional histopathology, the focus is on preservation of the tissue morphology which is usually achieved by fixation in formalin and embedding in paraffin wax. Formalin cross-links and hence damages proteins; the process also promotes RNA fragmentation. While techniques which do not need full-length sequence, such as *in situ* hybridisation, or even RT-PCR, can succeed in such material, as shown in Section 5.1.1.8, formalin-fixed paraffin-embedded tissue does not represent a suitable substrate for the more stringent methods requiring intact RNA, such as SAGE or DNA microarray analysis. For these, freezing tissue works well but it compromises the morphology needed for accurate microscopic diagnosis. A possible compromise is emerging in the form of alcohol-based fixatives which may preserve both morphology and macromolecules including mRNA (Liotta *et al.* 2000).

The method of tissue or cell selection provides a further technical challenge (Liotta *et al.* 2000), with options including bulk and microdissected tissue. Bulk tissue provides larger quantities of RNA and is technically easier to produce. It was my choice for this project, because at the time SAGE required large quantities of input RNA (see Section 3.1.2.1). The alternative is microdissected tissue which comprises a more select but smaller population of the exact cells desired, so that, for example, epithelial cells can be separated from stromal components. The most common method used currently is laser-assisted microdissection which permits the precise removal of pure cell populations from morphologically preserved tissue sections, in a rapid and practical manner (Goldsworthy *et al.* 1999; Sirivatanauksorn *et al.* 1999). Only small quantities of RNA are produced, but this problem can be solved by subsequent amplification. Microdissection is especially useful for studying premalignant disease, for example intestinal metaplasia or dysplasia, which by its nature is present in smaller amounts and is often focal and only identifiable histologically, not by gross examination (Liotta *et al.* 2000).

The ethical issues surrounding research on human tissues have become a minefield over recent years. In the UK there has been a recent sudden restriction in the availability of this basic resource (Furness 2001). This was first triggered by adverse publicity about inappropriate retention of whole organs at paediatric post-mortems, but the reaction has spread also to limit the use of tissues left over after routine diagnostic assessment of surgical resections, diagnostic biopsies, and even blood samples. Unquestionably, maintaining the autonomy of and obtaining informed consent from individual patients is important and necessary. However, the result of the media outcry has been that the regulatory authorities have swiftly imposed rather restrictive, oversimplified guidelines, which are bureaucratic and are now hampering clinical and translational cancer research.

Once the samples have been obtained and studied, the resulting large-scale gene expression profiles comprise a vast amount of data. Its storage, analysis, interpretation and use present a further challenge to scientists and clinicians. Bioinformatics is thus an expanding and important field of research, which is now the target of special funding.

In order to establish the diagnostic, prognostic and therapeutic importance of each of the emerging cancer gene candidates, analysis of hundreds of specimens from patients in different stages of disease will be needed. Selection of appropriate patient groups will therefore be vital. Clinical trials will need to include tissue collection as a key component thereby allowing current and future treatments to be evaluated by molecular methods. Such large-scale analysis of small numbers of targets in many tumour tissue samples is facilitated by the use of tissue arrays, in which around 300 cylindrical tissue cores of 0.6 mm diameter can be distributed in a single standard 2 x 3 cm wax block (Kononen *et al.* 1998). Parallel *in situ* detection of RNA and protein targets in each specimen is enabled and consecutive sections allow the rapid analysis of at least a hundred molecular markers in the same set of specimens.

Transfer to the real world of patient treatment will require the development of straightforward techniques for use in the clinical laboratory. Ideally these methods will involve a limited selection of predictive classifiers and so be suitable for high-throughput relatively automated analysis, in a manner akin to current immunohistochemical assessment of breast cancer. In this way, the exciting targets identified through large-scale tumour profiling can be translated to routine clinical diagnostic and prognostic use, resulting in improved care of cancer patients through individually tailored treatments.

7 Conclusion (This conclusion is an amalgam of those from chapters 3, 4 and 5.)

This thesis describes: the first global profiles of gene expression in gastric carcinoma and normal stomach created using serial analysis of gene expression (SAGE); and the resulting identification and subsequent characterisation of a new gene which was highly and specifically expressed in the stomach and which is likely to play an important role in normal gastric function.

First, the SAGE technique of large-scale mRNA expression profiling had to be established locally. Although the initial effort involved in setting up SAGE was considerable, requiring time, technical expertise and large amounts of high quality input RNA, the result has been a robust and reproducible technology, and the experience of establishing SAGE has given me an in-depth training in molecular biology and bioinformatics. SAGE produces libraries which are extremely valuable, providing data which are truly comprehensive and quantitative, and which enable the identification of novel genes, as follows.

SAGE was used to produce two libraries of gastric adenocarcinoma of distal, intestinal type, and one library of normal gastric antrum. Numerous transcripts have been identified which are: highly expressed; differentially expressed between normal and tumour stomach; or gastric-specific by comparison with normal and tumour breast, colon, ovary, pancreas, prostate and mesothelium. Selected genes have been validated in a wider panel of 19 gastro-intestinal tissues by Northern blotting and immunohistochemistry.

The overall statistics of the three SAGE libraries agreed with previous publications, with 1% of genes being differentially expressed between gastric carcinoma and normal stomach. The SAGE profiles were corroborated by the validation studies. The most abundant transcripts included ribosomal and mitochondrial proteins, of which most were up-regulated in the tumours, as were other widely expressed genes including transcription factors, growth factors, and genes involved in signal transduction, protein turnover and cell invasion. This pattern is similar to other cancers, which supports the existence of common molecular targets for diagnosis and therapy. Most genes which were abundant or more highly expressed in normal stomach play a role in normal gastric function, including gastrin, lysozyme, mucins, trefoil factors and pepsinogens, of which some (up to 0.65%) were gastric-specific by comparison with other normal glandular tissues.

The molecular anatomy of the tumours correlated with their morphology. The expression profile of the well differentiated gastric carcinoma more closely resembled normal stomach than did the moderately differentiated tumour. Some genes up-regulated in the gastric carcinomas indicated the acquisition of an intestinal phenotype, which has long been recognised histologically. Some genes associated with normal gastric antrum were also abundant in, and had previously been proposed as specific markers of, pancreatic carcinoma. These expression profiles, unexpected in some cases, together with tumour heterogeneity, emphasise the need for candidates to be tested, and ideally localised, in a wide range of normal and diseased cells and tissues.

New candidate genes have been identified. Some transcripts had previously been characterised minimally or not at all in the stomach. Many SAGE tags with interesting expression patterns, some tumour-associated, lacked matching genes and await further characterisation and functional studies. Of the novel genes, one was highly and specifically expressed in normal stomach, at levels similar to gastrin, MUC5 and the trefoil factors, which are important in normal gastric physiology. The gene is expressed in native or metaplastic gastric surface and foveolar epithelium and hence has been named foveolin. Although it appears to be gastric-predominant, foveolin may also be expressed at low levels in the uterus, although evidence for this is conflicting. Foveolin is not expressed in gastric adenocarcinomas or cell lines.

Foveolin mRNAs and their predicted proteins have been structurally characterised from human, mouse and cow, and genomic sequences from the human and mouse have also been obtained. The mRNAs are around 700-800 bp in length. The genes contain six exons which extend over around 6 kb. In the human, the gene maps to chromosome 2. The proteins are around 186 amino acids in length and just over 20 kDa in molecular weight. The sequences are highly homologous between species, again suggesting that it plays an important role in the stomach. Foveolin lacks homology to any known protein sequence or motif, so its function can only be postulated on the basis of its mucosal location: it could involve mucosal protection and possibly defence against gastric acid. This is supported by the fact that foveolin's primary sequence contains a signal peptide, indicating that the protein is extracellular, either located on the cell surface or secreted. Its absence from gastric carcinomas suggests that it may act as a functional gastric tumour suppressor gene, analogous to pS2 (trefoil factor 1): pS2 is highly expressed in normal stomach and down-regulated in tumours, and pS2 knock-out mice develop gastric dysplasia then carcinoma. Further characterisation of foveolin's biological role is thus required and likely to be worthwhile.

These molecular portraits increase our knowledge about the genes involved in normal gastric function and in malignant change in the stomach, and provide a catalogue of candidates from which to develop markers for better diagnosis and therapy of gastric carcinoma.

8 References

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno and et al. (1991). "Complementary DNA sequencing: expressed sequence tags and human genome project." Science **252**(5013): 1651-6.

Adams, M. D., A. R. Kerlavage, F. R.D. and e. al (1995). "Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence." Nature **377 Supp**: 3-17.

Ahrendt, S. A., S. Halachmi, J. T. Chow, L. Wu, N. Halachmi, S. C. Yang, S. Wehage, J. Jen and D. Sidransky (1999). "Rapid p53 sequence analysis in primary lung cancer using an oligonucleotide probe array." Proc Natl Acad Sci U S A **96**(13): 7382-7.

Akhurst, R. J. and R. Derynck (2001). "TGF-beta signaling in cancer--a double-edged sword." Trends Cell Biol **11**(11): S44-51.

Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, L. M. Staudt and et al. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." Nature **403**(6769): 503-11.

Anbazhagan, R., T. Tihan, D. M. Bornman, J. C. Johnston, J. H. Saltz, A. Weigering, S. Piantadosi and E. Gabrielson (1999). "Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles." Cancer Res **59**(20): 5119-22.

Angelastro, J. M., L. P. Klimaschewski and O. V. Vitolo (2000). "Improved NlaIII digestion of PAGE-purified 102 bp ditags by addition of a single purification step in both the SAGE and microSAGE protocols." Nucleic Acids Res **28**(12): E62.

Argani, P., C. Rosty, R. E. Reiter, R. E. Wilentz, S. R. Murugesan, S. D. Leach, B. Ryu, H. G. Skinner, M. Goggins, E. M. Jaffee, C. J. Yeo, J. L. Cameron, S. E. Kern and R. H. Hruban (2001). "Discovery of New Markers of Cancer through Serial Analysis of Gene Expression: Prostate Stem Cell Antigen Is Overexpressed in Pancreatic Adenocarcinoma." Cancer Res **61**(11): 4320-4324.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and

G. Sherlock (2000). “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.” Nat Genet **25**(1): 25-9.

Audic, S. and J.-M. Claverie (1997). “The Significance of Digital Gene Expression Profiles.” Genome Res. **7**: 986-995.

Bahrenberg, G., A. Brauers, H. G. Joost and G. Jakse (2000). “Reduced expression of PSCA, a member of the LY-6 family of cell surface antigens, in bladder, esophagus, and stomach tumors.” Biochemical and Biophysical Research Communications **275**(3): 783-788.

Bains, W. (1996). “Virtually sequenced: The next genomic generation.” Nature Biotechnology **14**: 711-713.

Barranco, S. C., C. M. Townsend, Jr., C. Casartelli, B. G. Macik, N. L. Burger, W. R. Boerwinkle and W. K. Gourley (1983). “Establishment and characterization of an in vitro model system for human adenocarcinoma of the stomach.” Cancer Res **43**(4): 1703-9.

Barrans, S. L., I. Carter, R. G. Owen, F. E. Davies, R. D. Patmore, A. P. Haynes, G. J. Morgan and A. S. Jack (2002). “Germinal center phenotype and bcl-2 expression combined with the International Prognostic Index improves patient risk stratification in diffuse large B-cell lymphoma.” Blood **99**(4): 1136-43.

Beck, F., F. Tata and K. Chawengsaksophak (2000). “Homeobox genes and gut development.” Bioessays **22**(5): 431-41.

Bell, D. A. and D. M. DeMarini (1991). “Excessive cycling converts PCR products to random-length higher molecular weight fragments.” Nucleic Acids Res **19**(18): 5079.

Berner, A. and J. M. Nesland (1991). “Endocrine profile in gastric carcinomas. An immunohistochemical study.” Histol Histopathol **6**(3): 317-23.

Bevan, S. and R. S. Houlston (1999). “Genetic predisposition to gastric cancer.” Qjm **92**(1): 5-10.

Blaser, M. J. (1998). “Helicobacter pylori and gastric diseases.” Bmj **316**(7143): 1507-10.

Blok, P., M. E. Craanen, G. J. A. Offerhaus and G. N. J. Tytgat (1997). “Gastric carcinoma: clinical, pathogenic and molecular aspects.” Q J Med **90**: 735-749.

Blume-Jensen, P. and T. Hunter (2001). “Oncogenic kinase signalling.” Nature **411**(6835): 355-65.

Boman, F., M. P. Buisine, A. Wacrenier, D. Querleu, J. P. Aubert and N. Porchet (2001). “Mucin gene transcripts in benign and borderline mucinous tumours of the ovary: an in situ hybridization study.” J Pathol **193**(3): 339-44.

Breitwieser, G. E., J. C. McLenithan, J. F. Cortese, J. M. Shields, M. M. Oliva, J. L. Majewski, C. E. Machamer and V. W. Yang (1997). "Colonic epithelium-enriched protein A4 is a proteolipid that exhibits ion channel characteristics." Am J Physiol **272**(3 Pt 1): C957-65.

Brent, R. (2000). "Genomic biology." Cell **100**(1): 169-83.

Brodeur, G. M., J. Pritchard, F. Berthold, N. L. Carlsen, V. Castel, R. P. Castelberry, B. De Bernardi, A. E. Evans, M. Favrot, F. Hedborg and et al. (1993). "Revisions of the international criteria for neuroblastoma diagnosis, staging, and response to treatment." J Clin Oncol **11**(8): 1466-77.

Brown, P. O. and D. Botstein (1999). "Exploring the new world of the genome with DNA microarrays." Nat Genet **21**(1 Suppl): 33-7.

Calam, J. and J. H. Baron (2001). "ABC of the upper gastrointestinal tract: Pathophysiology of duodenal and gastric ulcer and gastric cancer." Bmj **323**(7319): 980-2.

Caldas, C. and S. A. Aparicio (2002). "The molecular outlook." Nature **415**(6871): 484-5.

Campaign, Cancer Research. (1995). Stomach Cancer - UK.

Campaign, Cancer Research. (2001). CRC CancerStats: Mortality - UK.

Carulli, J. P., M. Artinger, P. M. Swain, C. D. Root, L. Chee, C. Tulig, J. Guerin, M. Osborne, G. Stein, J. Lian and P. T. Lomedico (1998). "High throughput analysis of differential gene expression." Journal of Cellular Biochemistry. Issue SUPPL **30**(31): 286-296.

Chan, A. O., J. M. Luk, W. M. Hui and S. K. Lam (1999). "Molecular biology of gastric carcinoma: from laboratory to bedside." J Gastroenterol Hepatol **14**(12): 1150-60.

Chang, K. and I. Pastan (1996). "Molecular cloning of mesothelin, a differentiation antigen present on mesothelium, mesotheliomas, and ovarian cancers." Proc Natl Acad Sci U S A **93**(1): 136-40.

Chen, J. J., J. D. Rowley and S. M. Wang (2000). "Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification." Proceedings of the National Academy of Sciences of the United States of America **97**(1): 349-353.

Corfield, A. P., N. Myerscough, R. Longman, P. Sylvester, S. Arul and M. Pignatelli (2000). "Mucins and mucosal protection in the gastrointestinal tract: new prospects for mucins in the pathology of gastrointestinal disease." Gut **47**(4): 589-94.

Correa, P. (1988). "A human model of gastric carcinogenesis." Cancer Research **48**(13): 3554-3560.

Correa, P. (1992). "Human gastric carcinogenesis: A multistep and multifactorial process - First American Cancer Society Award lecture on cancer epidemiology and prevention." Cancer Research **52**(24): 6735-6740.

Cotran, R. S., V. Kumar, S. L. Robbins and F. J. Schoen (1994). Robbins Pathologic Basis of Disease. Philadelphia, W.B. Saunders Company.

Datson, N. A., J. vanderPerkdeJong, M. P. vandenBerg, E. R. deKloet and E. Vreugdenhil (1999). "MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue." Nucleic Acids Research **27**(5): 1300-1307.

de Waard, V., B. M. M. van den Berg, J. Veken, R. Schultz-Heienbrok, H. Pannekoek and A. J. van Zonneveld (1999). "Serial analysis of gene expression to assess the endothelial cell response to an atherogenic stimulus." Gene **226**(1): 1-8.

Doglioni, C., L. Laurino, A. P. Dei Tos, M. De Boni, G. Franzin, P. Braidotti and G. Viale (1993). "Pancreatic (acinar) metaplasia of the gastric mucosa. Histology, ultrastructure, immunocytochemistry, and clinicopathologic correlations of 101 cases." Am J Surg Pathol **17**(11): 1134-43.

Ebert, M. P., T. Gunther, J. Hoffmann, J. Yu, S. Miehlke, H. U. Schulz, A. Roessner, M. Korc and P. Malfertheiner (2000). "Expression of metallothionein II in intestinal metaplasia, dysplasia, and gastric cancer." Cancer Res **60**(7): 1995-2001.

Eddy, S. R. (1999). "Noncoding RNA genes." Curr Opin Genet Dev **9**(6): 695-9.

Eisenhauer, E. A. (2001). "From the molecule to the clinic--inhibiting HER2 to treat breast cancer." N Engl J Med **344**(11): 841-2.

El-Omar, E. M., M. Carrington, W. H. Chow, K. E. McColl, J. H. Bream, H. A. Young, J. Herrera, J. Lissowska, C. C. Yuan, N. Rothman, G. Lanyon, M. Martin, J. F. Fraumeni, Jr. and C. S. Rabkin (2000). "Interleukin-1 polymorphisms associated with increased risk of gastric cancer." Nature **404**(6776): 398-402.

El-Omar, E. M., K. Oien, A. El-Nujumi, D. Gillen, A. Wirz, S. Dahill, C. Williams, J. E. Ardill and K. E. McColl (1997). "Helicobacter pylori infection and chronic gastric acid hyposecretion." Gastroenterology **113**(1): 15-24.

El-Omar, E. M., K. Oien, L. S. Murray, A. El-Nujumi, A. Wirz, D. Gillen, C. Williams, G. Fullarton and K. E. McColl (2000). "Increased prevalence of precancerous changes in relatives of gastric cancer patients: critical role of H. pylori." Gastroenterology **118**(1): 22-30.

El-Rifai, W., H. F. Frierson, Jr., J. C. Harper, S. M. Powell and S. Knuutila (2001). "Expression profiling of gastric adenocarcinoma using cDNA array." Int J Cancer **92**(6): 832-8.

Emmert-Buck, M. R., R. L. Strausberg, D. B. Krizman, M. F. Bonaldo, R. F. Bonner, D. G. Bostwick, M. R. Brown, K. H. Buetow, R. F. Chuaqui, K. A. Cole, P. H. Duray, C. R. Englert, J. W. Gillespie, S. Greenhut, L. Grouse, L. W. Hillier, K. S. Katz, R. D. Klausner, V. Kuznetzov, A. E. Lash, G. Lennon, W. M. Linehan, L. A. Liotta, M. A. Marra, P. J. Munson, D. K. Ornstein, V. V. Prabhu, C. Prange, G. D. Schuler, M. B. Soares, C. M. Tolstoshev, C. D. Vocke and R. H. Waterston (2000). "Molecular profiling of clinical tissue specimens: feasibility and applications." Am J Pathol **156**(4): 1109-15.

Ewing, B. and P. Green (2000). "Analysis of expressed sequence tags indicates 35,000 human genes." Nat Genet **25**(2): 232-4.

Fearon, E. R. and B. Vogelstein (1990). "A genetic model for colorectal tumorigenesis." Cell **61**(5): 759-767.

Friedl, A., S. P. Stoesz, P. Buckley and M. N. Gould (1999). "Neutrophil gelatinase-associated lipocalin in normal and neoplastic human tissues. Cell type-specific pattern of expression." Histochem J **31**(7): 433-41.

Fuchs, C. S. and R. J. Mayer (1995). "Medical progress: Gastric carcinoma." New England Journal of Medicine **333**(1): 32-41.

Furness, P. N. (2001). "Research using human tissues - a crisis of supply?" J Pathol **195**(3): 277-84.

Gibson, P. R., O. Rosella, A. J. Wilson, J. M. Mariadason, K. Rickard, K. Byron and D. H. Barkla (1999). "Colonic epithelial cell activation and the paradoxical effects of butyrate." Carcinogenesis **20**(4): 539-44.

Goldsworthy, S. M., P. S. Stockton, C. S. Trempus, J. F. Foley and R. R. Maronpot (1999). "Effects of fixation on RNA extraction and amplification from laser capture microdissected tissue." Mol Carcinog **25**(2): 86-91.

Green, K. J. and C. A. Gaudry (2000). "Are desmosomes more than tethers for intermediate filaments?" Nat Rev Mol Cell Biol **1**(3): 208-16.

Group, T. E. S. (1993). "An international association between Helicobacter pylori infection and gastric cancer." Lancet **341**(8857): 1359-1362.

Guilford, P., J. Hopkins, J. Harraway, M. McLeod, N. McLeod, P. Harawira, H. Taite, R. Scoular, A. Miller and A. E. Reeve (1998). "E-cadherin germline mutations in familial gastric cancer." Nature **392**(6674): 402-405.

Gygi, S. P., Y. Rochon, B. R. Franza and R. Aebersold (1999). "Correlation between protein and mRNA abundance in yeast." Molecular and Cellular Biology **19**(3): 1720-1730.

Hanahan, D. and R. A. Weinberg (2000). "The hallmarks of cancer." Cell **100**(1): 57-70.

Harris, A. and J. J. Misiewicz (2001). "ABC of the upper gastrointestinal tract. Management of Helicobacter pylori infection." Bmj **323**(7320): 1047-50.

Harris, R. M., R. Picton, S. Singh and R. H. Waring (2000). "Activity of phenolsulfotransferases in the human gastrointestinal tract." Life Sci **67**(17): 2051-7.

Harris, V., A. L. Sandridge, R. J. Black, D. H. Brewster and A. Gould (1998). Cancer Registration Statistics Scotland 1986 - 1995. Edinburgh, ISD Scotland Publications.

He, T. C., T. A. Chan, B. Vogelstein and K. W. Kinzler (1999). "PPARdelta is an APC-regulated target of nonsteroidal anti-inflammatory drugs." Cell **99**(3): 335-45.

He, T. C., A. B. Sparks, C. Rago, H. Hermeking, L. Zawel, L. T. da Costa, P. J. Morin, B. Vogelstein and K. W. Kinzler (1998). "Identification of c-MYC as a target of the APC pathway." Science **281**(5382): 1509-1512.

Henry, J. A., M. K. Bennett, N. H. Piggott, D. L. Levett, F. E. May and B. R. Westley (1991). "Expression of the pNR-2/pS2 protein in diverse human epithelial tumours." Br J Cancer **64**(4): 677-82.

Hibi, K., Q. Liu, G. A. Beaudry, S. L. Madden, W. H. Westra, S. L. Wehage, S. C. Yang, R. F. Heitmiller, A. H. Bertelsen, D. Sidransky and J. Jen (1998). "Serial analysis of gene expression in non-small cell lung cancer." Cancer Research **58**(24): 5690-5694.

Hippo, Y., M. Yashiro, M. Ishii, H. Taniguchi, S. Tsutsumi, K. Hirakawa, T. Kodama and H. Aburatani (2001). "Differential gene expression profiles of scirrhous gastric cancer cells with high metastatic potential to peritoneum or lymph nodes." Cancer Res **61**(3): 889-95.

Hough, C. D., C. A. Sherman-Baust, E. S. Pizer, F. J. Montz, D. D. Im, N. B. Rosenshein, K. R. Cho, G. J. Riggins and P. J. Morin (2000). "Large-scale serial analysis of gene expression reveals genes differentially expressed in ovarian cancer." Cancer Research **60**(22): 6281-6287.

Hubank, M. and D. G. Schatz (1994). "Identifying differences in mRNA expression by representational difference analysis of cDNA." Nucleic Acids Res **22**(25): 5640-8.

Huff, T., C. S. Muller, A. M. Otto, R. Netzker and E. Hannappel (2001). "beta-Thymosins, small acidic peptides with multiple functions." Int J Biochem Cell Biol **33**(3): 205-20.

Huntsman, D. G., F. Carneiro, F. R. Lewis, P. M. MacLeod, A. Hayashi, K. G. Monaghan, R. Maung, R. Seruca, C. E. Jackson and C. Caldas (2001). "Early gastric cancer in young, asymptomatic carriers of germ-line E-cadherin mutations." N Engl J Med **344**(25): 1904-9.

Hurst, A. F. and M. J. Stewart (1929). Gastric and Duodenal Ulcer. Oxford, Oxford University Press.

Isaacson, P. (1982). "Immunoperoxidase study of the secretory immunoglobulin system and lysozyme in normal and diseased gastric mucosa." Gut **23**(7): 578-88.

Jakowlew, S. B., R. Breathnach, J. M. Jeltsch, P. Masiakowski and P. Chambon (1984). "Sequence of the pS2 mRNA induced by estrogen in the human breast cancer cell line MCF-7." Nucleic Acids Res **12**(6): 2861-78.

Janke, J., K. Schluter, B. Jandrig, M. Theile, K. Kolble, W. Arnold, E. Grinstein, A. Schwartz, L. Estevez-Schwarz, P. M. Schlag, B. M. Jockusch and S. Scherneck (2000). "Suppression of tumorigenicity in breast cancer cells by the microfilament protein profilin 1." J Exp Med **191**(10): 1675-86.

Jung, M. H., S. C. Kim, G. A. Jeon, S. H. Kim, Y. Kim, K. S. Choi, S. I. Park, M. K. Joe and K. Kimm (2000). "Identification of differentially expressed genes in normal and tumor human gastric tissue." Genomics **69**(3): 281-6.

Kal, A. J., A. J. van Zonneveld, V. Benes, M. van den Berg, M. G. Koerkamp, K. Albermann, N. Strack, J. M. Ruijter, A. Richter, B. Dujon, W. Ansorge and H. F. Tabak (1999). "Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources." Molecular Biology of the Cell **10**(6): 1859-1872.

Kawai, J., A. Shinagawa, K. Shibata, M. Yoshino, M. Itoh, Y. Ishii, T. Arakawa, A. Hara, Y. Fukunishi, H. Konno, J. Adachi, S. Fukuda, K. Aizawa, M. Izawa, K. Nishi, H. Kiyosawa, S. Kondo, I. Yamanaka, T. Saito, Y. Okazaki, T. Gojobori, H. Bono, T. Kasukawa, R. Saito, K. Kadota, H. A. Matsuda, M. Ashburner, S. Batalov, T. Casavant, W. Fleischmann, T. Gaasterland, C. Gissi, B. King, H. Kochiwa, P. Kuehl, S. Lewis, Y. Matsuo, I. Nikaido, G. Pesole, J. Quackenbush, L. M. Schriml, F. Staubli, R. Suzuki, M. Tomita, L. Wagner, T. Washio, K. Sakai, T. Okido, M. Furuno, H. Aono, R. Baldarelli, G. Barsh, J. Blake, D. Boffelli, N. Bojunga, P. Carninci, M. F. de Bonaldo, M. J. Brownstein, C. Bult, C. Fletcher, M. Fujita, M. Gariboldi, S. Gustincich, D. Hill, M. Hofmann, D. A. Hume, M. Kamiya, N. H. Lee, P. Lyons, L. Marchionni, J. Mashima, J. Mazzarelli, P. Mombaerts, P. Nordone, B. Ring, M. Ringwald, I. Rodriguez, N. Sakamoto, H. Sasaki, K. Sato, C. Schonbach, T. Seya, Y. Shibata, K. F. Storch, H. Suzuki, K. Toyo-oka, K. H. Wang, C. Weitz, C. Whittaker, L. Wilming, A. Wynshaw-Boris, K. Yoshida, Y.

Hasegawa, H. Kawaji, S. Kohtsuki and Y. Hayashizaki (2001). "Functional annotation of a full-length mouse cDNA collection." Nature **409**(6821): 685-90.

Kenzelmann, M. and K. Muhlemann (1999). "Substantially enhanced cloning efficiency of SAGE (Serial Analysis of Gene Expression) by adding a heating step to the original protocol." Nucleic Acids Research **27**(3): 917-918.

Khan, J., J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson and P. S. Meltzer (2001). "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." Nat Med **7**(6): 673-9.

Klein Kranenbarg, E., J. Hermans, J. H. van Krieken and C. J. van de Velde (2001). "Evaluation of the 5th edition of the TNM classification for gastric cancer: improved prognostic value." Br J Cancer **84**(1): 64-71.

Konishi, N., K. Matsumoto, Y. Hiasa, Y. Kitahori, I. Hayashi and H. Matsuda (1995). "Tissue and serum pepsinogen I and II in gastric cancer identified using immunohistochemistry and rapid ELISA." J Clin Pathol **48**(4): 364-7.

Kononen, J., L. Bubendorf, A. Kallioniemi, M. Barlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter and O. P. Kallioniemi (1998). "Tissue microarrays for high-throughput molecular profiling of tumor specimens." Nat Med **4**(7): 844-7.

Kozak, M. (1996). "Interpreting cDNA sequences: some insights from studies on translation." Mamm Genome **7**(8): 563-74.

Kushima, R., R. Manabe, T. Hattori and F. Borchard (1999). "Histogenesis of gastric foveolar metaplasia following duodenal ulcer: a definite reparative lineage of Brunner's gland." Histopathology **35**(1): 38-43.

Kuska, B. (1996). "Cancer Genome Anatomy Project Set for Take-Off (News)." J Natl Can Inst **88**(24): 1801-1803.

Lal, A., A. E. Lash, S. F. Altschul, V. Velculescu, L. Zhang, R. E. McLendon, M. A. Marra, C. Prange, P. J. Morin, K. Polyak, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, R. L. Strausberg and G. J. Riggins (1999). "A public database for gene expression in human cancers." Cancer Research **59**(21): 5403-5407.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D.

Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan and J. Szustakowki (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Larsson, M., S. Stahl, M. Uhlen and A. Wennborg (2000). "Expression profile viewer (ExProView): A software tool for transcriptome analysis." Genomics **63**(3): 341-353.

Laurén, P. (1965). "The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. An attempt at a histoclinical classification." Acta Pathol Microbiol Scand **64**: 31-49.

Lee, S. W., C. Tomasetto and R. Sager (1991). "Positive selection of candidate tumor-suppressor genes by subtractive hybridization." Proc Natl Acad Sci U S A **88**(7): 2825-9.

Lefebvre, O., M. P. Chenard, R. Masson, J. Linares, A. Dierich, M. LeMeur, C. Wendling, C. Tomasetto, P. Chambon and M. C. Rio (1996). "Gastric mucosa abnormalities and tumorigenesis in mice lacking the pS2 trefoil protein." Science **274**(5285): 259-262.

Lennon, G., C. Auffray, M. Polymeropoulos and M. B. Soares (1996). "The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression." Genomics **33**(1): 151-2.

Lewin, K. J. and H. D. Appelman (1996). Tumors of the Esophagus and Stomach. Washington, D.C., Armed Forces Institute of Pathology.

Liang, P. and A. B. Pardee (1992). "Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction." Science **257**(5072): 967-71.

Liotta, L. and E. Petricoin (2000). "Molecular profiling of human cancer." Nat Rev Genet **1**(1): 48-56.

Lockhart, D. J. and E. A. Winzeler (2000). "Genomics, gene expression and DNA arrays." Nature **405**(6788): 827-836.

Logan, R. P. and M. M. Walker (2001). "ABC of the upper gastrointestinal tract: Epidemiology and diagnosis of Helicobacter pylori infection." Bmj **323**(7318): 920-2.

Luqmani, Y., C. Bennett, I. Paterson, C. M. Corbishley, M. C. Rio, P. Chambon and G. Ryall (1989). "Expression of the pS2 gene in normal, benign and neoplastic human stomach." Int J Cancer **44**(5): 806-12.

Ma, T. and A. S. Verkman (1999). "Aquaporin water channels in gastrointestinal physiology." J Physiol **517**(Pt 2): 317-26.

Macdonald, J. S., S. R. Smalley, J. Benedetti, S. A. Hundahl, N. C. Estes, G. N. Stemmermann, D. G. Haller, J. A. Ajani, L. L. Gunderson, J. M. Jessup and J. A. Martenson (2001). "Chemoradiotherapy after surgery compared with surgery alone for adenocarcinoma of the stomach or gastroesophageal junction." N Engl J Med **345**(10): 725-30.

Machado, J. C., F. Carneiro, P. Ribeiro, N. Blin and M. Sobrinho-Simoes (1996). "pS2 protein expression in gastric carcinoma. An immunohistochemical and immunoradiometric study." Eur J Cancer **32A**(9): 1585-90.

Machado, J. C., A. M. Nogueira, F. Carneiro, C. A. Reis and M. Sobrinho-Simoes (2000). "Gastric carcinoma exhibits distinct types of cell differentiation: an immunohistochemical study of trefoil peptides (TFF1 and TFF2) and mucins (MUC1, MUC2, MUC5AC, and MUC6)." J Pathol **190**(4): 437-43.

Madden, S. L., E. A. Galella, J. Zhu, A. H. Bertelsen and G. A. Beaudry (1997). "SAGE transcript profiles for p53-dependent growth regulation." Oncogene **15**(9): 1079-1085.

Madden, S. L., C. J. Wang and G. Landes (2000). "Serial analysis of gene expression: from gene discovery to target identification." Drug Discovery Today **5**(9): 415-425.

Man, M. Z., X. N. Wang and Y. X. Wang (2000). "POWER_SAGE: comparing statistical tests for SAGE experiments." Bioinformatics **16**(11): 953-959.

Margulies, E. H. and J. W. Innis (2000). "eSAGE: managing and analysing data generated with Serial Analysis of Gene Expression (SAGE)." Bioinformatics **16**(7): 650-651.

Margulies, E. H., S. L. Kardia and J. W. Innis (2001). "Identification and prevention of a GC content bias in SAGE libraries." Nucleic Acids Res **29**(12): E60-0.

Marshall, B. J. and J. R. Warren (1984). "Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration." Lancet **1**(8390): 1311-5.

Matsumura, H., S. Nirasawa and R. Terauchi (1999). "Transcript profiling in rice (Oryza sativa L.) seedlings using serial analysis of gene expression (SAGE)." Plant Journal **20**(6): 719-726.

Mauro, M. J., M. O'Dwyer, M. C. Heinrich and B. J. Druker (2002). "STI571: a paradigm of new agents for cancer therapeutics." J Clin Oncol **20**(1): 325-34.

McAleer, J. J. A. (2001). Gastrointestinal cancer. Oncology. S. R. A. J. and J. P. G. Oxford, Oxford University Press.

McNicol, A. M. and M. A. Farquharson (1997). "In situ hybridization and its diagnostic applications in pathology." J Pathol **182**(3): 250-61.

Menke-Pluymers, M., E. v. Drunen, K. Vissers, A. Mulder, H. Tilanus and A. Hagemeijer (1996). "Cytogenetic analysis of Barrett's mucosa and adenocarcinoma of the distal esophagus and cardia." Cancer Genet Cytogenet **90**(2): 109-117.

Mitani, M., Y. Kuwabara, H. Kawamura, A. Sato, K. Hattori and Y. Fujii (2000). "Significance of plasma thymosin alpha1 measurements in gastric cancer patients." World J Surg **24**(4): 455-8.

Mitelman, F., F. Mertens and B. Johansson (1997). "A breakpoint map of recurrent chromosomal rearrangements in human neoplasia." Nat Genet **15**(Spec No): 417-74.

Nacht, M., T. Dracheva, Y. Gao, T. Fujii, Y. Chen, A. Player, V. Akmaev, B. Cook, M. Dufault, M. Zhang, W. Zhang, M. Guo, J. Curran, S. Han, D. Sidransky, K. Buetow, S. L. Madden and J. Jen (2001). "Molecular characteristics of non-small cell lung cancer." Proc Natl Acad Sci U S A **98**(26): 15203-8.

Nacht, M., A. T. Ferguson, W. Zhang, J. M. Petroziello, B. P. Cook, Y. H. Gao, S. Maguire, D. Riley, G. Coppola, G. M. Landes, S. L. Madden and S. Sukumar (1999). "Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer." Cancer Research **59**(21): 5464-5470.

Naito, Y., I. Kino, K. Horiuchi and D. Fujimoto (1984). "Promotion of collagen production by human fibroblasts with gastric cancer cells in vitro." Virchows Arch B Cell Pathol Incl Mol Pathol **46**(1-2): 145-54.

Nakai, K. and M. Kanehisa (1992). "A knowledge base for predicting protein localization sites in eukaryotic cells." Genomics **14**(4): 897-911.

Neilson, L., A. Andalibi, D. Kang, C. Coutifaris, J. F. Strauss, J. A. L. Stanton and D. P. L. Green (2000). "Molecular phenotype of the human oocyte by PCR-SAGE [Full text delivery]." Genomics **63**(1): 13-24.

Newton, J. L., A. Allen, B. R. Westley and F. E. May (2000). "The human trefoil peptide, TFF1, is present in different molecular forms that are intimately associated with mucus in normal stomach." Gut **46**(3): 312-20.

Norton, J. D. (2000). "ID helix-loop-helix proteins in cell growth, differentiation and tumorigenesis." J Cell Sci **113**(Pt 22): 3897-905.

Osborne, C. K. (1998). "Tamoxifen in the treatment of breast cancer." N Engl J Med **339**(22): 1609-18.

Osoegawa, K., A. G. Mammoser, C. Wu, E. Frengen, C. Zeng, J. J. Catanese and P. J. de Jong (2001). "A bacterial artificial chromosome library for sequencing the complete human genome." Genome Res **11**(3): 483-96.

Owen, D. A. (1986). "Normal histology of the stomach." Am J Surg Pathol **10**(1): 48-61.

Pandey, A. and M. Mann (2000). "Proteomics to study genes and genomes." Nature **405**(6788): 837-46.

Parkin, D. M. (2001). "Global cancer statistics in the year 2000." Lancet Oncol **2**: 533-43.

Peddanna, N., S. Holt and R. S. Verma (1995). "Genetics of gastric cancer." Anticancer Research **15**(5 B): 2055-2064.

Perou, C. M., T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A. L. Borresen-Dale, P. O. Brown and D. Botstein (2000). "Molecular portraits of human breast tumours." Nature **406**(6797): 747-52.

Pesole, G., G. Bernardi and C. Saccone (1999). "Isochore specificity of AUG initiator context of human genes." FEBS Lett **464**(1-2): 60-2.

Peters, D. G., A. B. Kassam, H. Yonas, E. H. O'Hare, R. E. Ferrell and A. M. Brufsky (1999). "Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite." Nucleic Acids Res **27**(24): e39.

Petricoin, E. F., A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn and L. A. Liotta (2002). "Use of proteomic patterns in serum to identify ovarian cancer." Lancet **359**(9306): 572-7.

Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray and D. G. Albertson (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays." Nat Genet **20**(2): 207-11.

Pollack, J. R., C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein and P. O. Brown (1999). "Genome-wide analysis of DNA copy-number changes using cDNA microarrays." Nat Genet **23**(1): 41-6.

Polyak, K. and G. J. Riggins (2001). "Gene discovery using the serial analysis of gene expression technique: implications for cancer research." J Clin Oncol **19**(11): 2948-58.

Polyak, K., Y. Xia, J. L. Zweler, K. W. Kinzler and B. Vogeistein (1997). "A model for p53-induced apoptosis." Nature **389**(6648): 300-305.

Powell, J. (1998). "Enhanced concatemer cloning - a modification to the SAGE (Serial Analysis of Gene Expression) technique." Nucleic Acids Research **26**(14): 3445-3446.

Pradet-Balade, B., F. Boulme, H. Beug, E. W. Mullner and J. A. Garcia-Sanz (2001). "Translation control: bridging the gap between genomics and proteomics?" Trends Biochem Sci **26**(4): 225-9.

Rockett, J. C., K. Larkin, S. J. Darnton, A. G. Morris and H. R. Matthews (1997). "Five newly established oesophageal carcinoma cell lines: phenotypic and immunological characterization." Br J Cancer **75**(2): 258-63.

Roukos, D. H. (2000). "Current status and future perspectives in gastric cancer management." Cancer Treat Rev **26**(4): 243-55.

Ryu, B., J. Jones, M. A. Hollingsworth, R. H. Hruban and S. E. Kern (2001). "Invasion-specific genes in malignancy: Serial analysis of gene expression comparisons of primary and passaged cancers." Cancer Research **61**(5): 1833-1838.

Salesiotis, A. N., C. K. Wang, C. D. Wang, A. Burger, H. Li and A. Seth (1995). "Identification of novel genes from stomach cancer cell lines by differential display." Cancer Lett **91**(1): 47-54.

Sambrook, J., E. F. Fritsch and T. Maniatis (1987). Molecular Cloning, Cold Spring Harbor Laboratory Press.

Schaer, J. C. and J. C. Reubi (1999). "High gastrin and cholecystokinin (CCK) gene expression in human neuronal, renal, and myogenic stem cell tumors: comparison with CCK-A and CCK-B receptor contents." J Clin Endocrinol Metab **84**(1): 233-9.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-70.

Schena, M., D. Shalon, R. Heller, A. Chai, P. O. Brown and R. W. Davis (1996). "Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes." Proceedings of the National Academy of Sciences of the United States of America **93**(20): 10614-10619.

Scheurle, D., M. P. DeYoung, D. M. Binninger, H. Page, M. Jahanzeb and R. Narayanan (2000). "Cancer gene discovery using digital differential display." Cancer Res **60**(15): 4037-43.

Schlemper, R. J., M. Itabashi, Y. Kato, K. J. Lewin, R. H. Riddell, T. Shimoda, P. Sipponen, M. Stolte, H. Watanabe, H. Takahashi and R. Fujita (1997). "Differences in diagnostic criteria for gastric carcinoma between Japanese and Western pathologists." Lancet **349**(9067): 1725-1729.

Schuler, G. D. (1997). "Pieces of the puzzle: expressed sequence tags and the catalog of human genes." J Mol Med **75**(10): 694-8.

Schwartz, A. L. and A. Ciechanover (1999). "The ubiquitin-proteasome pathway and pathogenesis of human diseases." Annu Rev Med **50**: 57-74.

Sekiguchi, M., K. Sakakibara and G. Fujii (1978). "Establishment of cultured cell lines derived from a human gastric carcinoma." Jpn J Exp Med **48**(1): 61-8.

Sessa, F., M. Bonato, B. Frigerio, C. Capella, E. Solcia, M. Prat, J. Bara and I. M. Samloff (1990). "Ductal cancers of the pancreas frequently express markers of gastrointestinal epithelial cells." Gastroenterology **98**(6): 1655-65.

Shimada, K. and J. A. Ajani (1999). "Adjuvant therapy for gastric carcinoma patients in the past 15 years: a review of western and oriental trials." Cancer **86**(9): 1657-68.

Sidransky, D. (1997). "Nucleic acid-based methods for the detection of cancer." Science **278**(5340): 1054-9.

Simpson, P. J. (1987). "An improved method for mRNA isolation and characterization of in vitro translation products by Western blotting." Gene **56**(2-3): 161-71.

Sirivatanauksorn, Y., R. Drury, T. CrnogoracJurcevic, V. Sirivatanauksorn and N. R. Lemoine (1999). "Laser-assisted microdissection: Applications in molecular pathology." Journal of Pathology **189**(2): 150-154.

Solcia, E., R. Fiocca, O. Luinetti, L. Villani, L. Padovan, D. Calistri, G. N. Ranzani, A. Chiaravalli and C. Capella (1996). "Intestinal and diffuse gastric cancers arise in a different background of Helicobacter pylori gastritis through different gene involvement." American Journal of Surgical Pathology **20**(SUPPL. 1): S8-S22.

Stollberg, J., J. Urschitz, Z. Urban and C. D. Boyd (2000). "A quantitative evaluation of SAGE." Genome Research **10**(8): 1241-1248.

Stolzenberg-Solomon, R. Z., M. J. Blaser, P. J. Limburg, G. Perez-Perez, P. R. Taylor, J. Virtamo and D. Albanes (2001). "Helicobacter pylori seropositivity as a risk factor for pancreatic cancer." J Natl Cancer Inst **93**(12): 937-41.

Strachan, T. and A. P. Read (1999). Human Molecular Genetics 2. Oxford, BIOS Scientific Publishers Ltd.

Strausberg, R. L., S. F. Greenhut, L. H. Grouse, C. F. Schaefer and K. H. Buetow (2001). "In silico analysis of cancer through the Cancer Genome Anatomy Project." Trends Cell Biol **11**(11): S66-71.

Tahara, E., H. Kuniyasu, W. Yasui and H. Yokozaki (1994). "Gene alterations in intestinal metaplasia and gastric cancer." European Journal of Gastroenterology & Hepatology, Supplement **6**(1): S97-S102.

Tahara, E., S. Semba and H. Tahara (1996). "Molecular biological observations in gastric cancer." Seminars in Oncology **23**(3): 307-315.

Tenti, P., A. Aguzzi, C. Riva, L. Usellini, R. Zappatore, J. Bara, I. M. Samloff and E. Solcia (1992). "Ovarian mucinous tumors frequently express markers of gastric, intestinal, and pancreatobiliary epithelial cells." Cancer **69**(8): 2131-42.

Tenti, P., S. Romagnoli, E. Silini, R. Zappatore, P. Giunta, G. Stella and L. Carnevali (1994). "Cervical adenocarcinomas express markers common to gastric, intestinal, and pancreatobiliary epithelial cells." Pathol Res Pract **190**(4): 342-9.

Tortora, G. J. and S. R. Grabowski (2000). Principles of Anatomy and Physiology. New York, John Wiley & Sons, Inc.

Tschmelitsch, J., M. R. Weiser and M. S. Karpeh (2000). "Modern staging in gastric cancer." Surg Oncol **9**(1): 23-30.

Uemura, N., S. Okamoto, S. Yamamoto, N. Matsumura, S. Yamaguchi, M. Yamakido, K. Taniyama, N. Sasaki and R. J. Schlemper (2001). "Helicobacter pylori infection and the development of gastric cancer." N Engl J Med **345**(11): 784-9.

Valle, J. W. (2001). "Adjuvant therapy for gastric cancer - has the standard changed?" Br J Cancer **84**(7): 875-7.

van den Berg, A., J. van der Leij and S. Poppema (1999). "Serial analysis of gene expression: rapid RT-PCR analysis of unknown SAGE tags." Nucleic Acids Res **27**(17): e17.

van Kampen, A. H. C., B. D. C. van Schaik, E. Pauws, E. M. C. Michiels, J. M. Ruijter, H. N. Caron, R. Versteeg, S. H. Heisterkamp, J. A. M. Leunissen, F. Baas and M. van der Mee (2000). "USAGE: a web-based approach towards the analysis of SAGE data." Bioinformatics **16**(10): 899-905.

van 't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature **415**(6871): 530-6.

Velculescu, V., L. Zhang, W. Zhou, J. Vogelstein, M. Basrai, J. Bassett DE, P. Hieter, B. Vogelstein and K. Kinzler (1997). "Characterization of the yeast transcriptome." Cell **88**(2): 243-51.

Velculescu, V. E., S. L. Madden, L. Zhang, A. E. Lash, J. Yu, C. Rago, A. Lal, C. J. Wang, G. A. Beaudry, K. M. Ciriello, B. P. Cook, M. R. Dufault, A. T. Ferguson, Y. H. Gao, T. C. He, H. Hermeking, S. K. Hiraldo, P. M. Hwang, M. A. Lopez, H. F. Luderer, B. Mathews, J. M. Petroziello, K. Polyak, L. Zawel, W. Zhang, X. M. Zhang, W. Zhou, F. G. Haluska, J. Jen, S. Sukumar, G. M. Landes, G. J. Riggins, B. Vogelstein and K. W. Kinzler (1999). "Analysis of human transcriptomes." Nature Genetics **23**(4): 387-388.

Velculescu, V. E., B. Vogelstein and K. W. Kinzler (2000). "Analysing uncharted transcriptomes with SAGE." Trends in Genetics **16**(10): 423-425.

Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler (1995). "Serial analysis of gene expression." Science **270**(5235): 484-487.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M.

Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Virlon, B., L. Cheval, J. M. Buhler, E. Billon, A. Doucet and J. M. Elalouf (1999). "Serial microanalysis of renal transcriptomes." Proceedings of the National Academy of Sciences of the United States of America **96**(26): 15286-15291.

Waghray, A., M. Schober, F. Feroze, F. Yao, J. Virgin and Y. Q. Chen (2001). "Identification of differentially expressed genes by serial analysis of gene expression in human prostate cancer." Cancer Res **61**(10): 4283-6.

Wang, D. G., J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, E. S. Lander and et al. (1998). "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome." Science **280**(5366): 1077-82.

Wang, G., M. Wang, W. You and H. Li (2001). "[Cloning and primary expression analyses of down-regulated cDNA fragment in human gastric cancer]." Zhonghua Yi Xue Yi Chuan Xue Za Zhi **18**(1): 43-7.

Wheeler, D. L., D. M. Church, A. E. Lash, D. D. Leipe, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, T. A. Tatusova, L. Wagner and B. A. Rapp (2001). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **29**(1): 11-6.

Williams, G. R. and N. A. Wright (1997). "Trefoil factor family domain peptides." Virchows Arch **431**(5): 299-304.

Wong, W. M., R. Poulsom and N. A. Wright (1999). "Trefoil peptides." Gut **44**(6): 890-5.

Wu, W. J., J. W. Erickson, R. Lin and R. A. Cerione (2000). "The gamma-subunit of the coatomer complex binds Cdc42 to mediate transformation." Nature **405**(6788): 800-4.

Wyllie, A. D. (1992). Growth and Neoplasia. Muir's Textbook of Pathology. R. N. M. MacSween and K. Whaley. London, Edward Arnold.

Yamamoto, M., T. Wakatsuki, A. Hada and A. Ryo (2001). "Use of serial analysis of gene expression (SAGE) technology." Journal of Immunological Methods **250**(1-2): 45-66.

Yarden, Y. (2001). "The EGFR family and its ligands in human cancer. signalling mechanisms and therapeutic opportunities." Eur J Cancer **37**(Suppl 4): S3-8.

Yasui, W., H. Yokozaki, F. Shimamoto, H. Tahara and E. Tahara (1999). "Molecular-pathological diagnosis of gastrointestinal tissues and its contribution to cancer histopathology." Pathol Int **49**(9): 763-74.

Ye, S. Q., L. Q. Zhang, F. Zheng, D. Virgil and P. O. Kwiterovich (2000). "MiniSAGE: Gene expression profiling using serial analysis of gene expression from 1 mu g total RNA." Analytical Biochemistry **287**(1): 144-152.

Yoshikawa, Y., H. Mukai, F. Hino, K. Asada and I. Kato (2000). "Isolation of two novel genes, down-regulated in gastric cancer." Jpn J Cancer Res **91**(5): 459-63.

Younes, Z. and D. A. Johnson (1997). "Molecular and genetic advances in gastrointestinal cancer: state of the art." Dig Dis **15**(4-5): 275-301.

Zamboni, G., A. Scarpa, G. Bogina, C. Iacono, C. Bassi, G. Talamini, F. Sessa, C. Capella, E. Solcia, F. Rickaert, G. M. Mariuzzi and G. Kloppel (1999). "Mucinous cystic tumors of the pancreas: clinicopathological features, prognosis, and relationship to other mucinous cystic tumors." Am J Surg Pathol **23**(4): 410-22.

Zhang, L., W. Zhou, V. E. Velculescu, S. E. Kern, R. H. Hruban, S. R. Hamilton, B. Vogelstein and K. W. Kinzler (1997). "Gene expression profiles in normal and cancer cells." Science **276**(5316): 1268-1272.

Zhou, W., S. N. Goodman, G. Galizia, E. Lieto, F. Ferraraccio, C. Pignatelli, C. A. Purdie, J. Piris, R. Morris, D. J. Harrison, P. B. Paty, A. Culliford, K. E. Romans, E. A. Montgomery, M. A. Choti, K. W. Kinzler and B. Vogelstein (2002). "Counting alleles to predict recurrence of early-stage colorectal cancers." Lancet **359**(9302): 219-25.