



Wang, Erxuan (2024) *Application of CUT&Tag to the mapping and analysis of VEZF1 binding sites in K562 cells*. MSc(R) thesis.

<https://theses.gla.ac.uk/84670/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>
research-enlighten@glasgow.ac.uk

Application of CUT&Tag to the mapping and analysis of VEZF1 binding sites in K562 cells

Erxuan Wang

Thesis submitted in fulfilment of the requirements for the
Degree of Master of Science

School of Cancer Sciences
College of Medical, Veterinary and Life Sciences
University of Glasgow

April 13, 2024

Abstract

VEZF1 is a highly conserved vertebrate transcription factor that has ubiquitous expression in vertebrates. VEZF1 is essential for the barrier activity of the chicken β globin HS4 insulator, where it prevents *de novo* DNA methylation. Knock-out of VEZF1 results in lethal haemorrhaging and edema in murine embryos, indicating a role for VEZF1 in the maintenance of vascular integrity.

In this study, the CUT&Tag method, previously reported to be an improvement on CUT&RUN and ChIP-seq, was utilised to map VEZF1 binding sites in K562 cells. The binding sites identified by CUT&Tag are similar to those previously identified using ChIP-seq, but the data generated from CUT&Tag shows significant advantages of lower background and higher peaks. Peak analysis indicates that most of the VEZF1 peaks are associated with promoters or enhancers, and there is strong co-localisation of the binding of VEZF1 and GATA2. VEZF1 tends to recognize and bind to GGGNGGGG motifs, but no difference is found between GGGNGGGG motifs discovered at VEZF1-associated promoters and enhancers. Furthermore, co-localisation of GATA2 does not affect the sequence of the GGGNGGGG motifs enriched at VEZF1 peaks. Sequence analysis also revealed other motifs that are recognized by a variety of transcription factors, which may act alongside VEZF1 to regulate gene expression in K562 cells.

Our study shows that CUT&Tag is an efficient method for mapping and analyzing the binding sites of transcription factors, and provides new opportunities for research on the interactions between VEZF1 and other transcription factors.

Table of Contents

Abstract	2
List of figures	5
List of tables	6
Acknowledgments	7
Author's Declaration	8
Abbreviation	9
Chapter 1 Introduction	12
1.1 Development of the vascular system	12
1.2 VEZF1	13
1.3 Chromatin binding by VEZF1	15
1.4 Transcription factors regulating endothelial cell differentiation	16
1.4.1 GATA2	16
1.4.2 CTCF	18
1.5 Epigenetics and histone modifications	19
1.5.1 H3K4me1	19
1.5.2 H3K4me3	20
1.5.3 H3K27ac	20
1.6 CUT&Tag	21
1.7 Aims of the project	23
1.8 Objectives	23
Chapter 2 Materials and Methods	25
2.1 Cell culture	25
2.1.1 Cell thawing	25
2.1.2 Culture conditions	25
2.1.3 Cell counting	25
2.1.4 Cell passaging	26
2.2 Nuclear extract preparation	26
2.2.1 Preparation of stock solutions and buffer solutions	26
2.2.2 Harvesting cells	26
2.2.3 Swelling the cells	27
2.2.4 Cell disruption	27
2.2.5 Salt extraction of nuclear proteins from chromatin	27
2.3 Bradford Assay	27
2.4 Western blotting	28
2.5 Immunofluorescence	29
2.6 CUT&Tag	30
2.6.1 Sample preparation	30
2.6.2 Sample sequencing	30
2.6.3 Next Generation Sequencing (NGS) data processing and analysis	31
2.7 Antibodies	34
Chapter 3 Results	35
3.1 Preparation and concentration detection of cell nuclear extracts	35
3.2 Western blotting	36
3.2.1 Testing the original stock of VEZF1 antibody (AGW-3642-old)	36

	4
3.2.2 Purification and verification of new VEZF1 antibodies	38
3.2.3 Verification and comparison of old and newly purified VEZF1 antibodies	38
3.2.4 Selection and concentration optimization of commercial antibodies	39
3.3 Immunofluorescence	41
3.4 Use of CUT&Tag methodology to map the distribution of VEZF1 and other factors in K562 cells	43
3.4.1 Preparation of CUT&Tag samples for next generation sequencing	43
3.4.2 Quality control of NGS data	43
A	44
3.4.3 Alignment to the genome and visualisation of CUT&Tag data	45
3.4.4 Investigation into duplicate reads	46
3.4.5 Peak finding	47
3.4.6 Enrichment of VEZF1, GATA2, CTCF and histone modifications at the <i>TAL1</i> locus	51
A	52
3.4.7 Comparison of CUT&Tag data with CHIP-seq data	52
3.4.8 Analysis of peaks	54
A	56
3.4.9 Analysis of DNA sequence motifs	57
Chapter 4 Discussion	62
References	68

List of figures

Figure 1.1	Stepwise development of vessels of the three circulations.....	13
Figure 1.2	Model of hemangiogenic mesoderm cell differentiation by regulation of transcription factors.....	16
Figure 1.3	In situ tethering for CUT&Tag chromatin profiling.....	22
Figure 3.1	Determination of the concentration of two nuclear extracts.....	36
Figure 3.2	Testing the activity and specificity of the stored VEZF1 antibody.....	37
Figure 3.3	Verification and comparison of activity and specificity of three VEZF1 antibodies.....	38
Figure 3.4	Verification and optimization of commercial antibodies.....	39
Figure 3.5	Test of the activity and specificity of antibodies by immunofluorescence.....	42
Figure 3.6	FastQC on raw data.....	44
Figure 3.7	FastQC after trimming.....	45
Figure 3.8	VEZF1 bigwig tracks (duplicates removed) with MACS2 at <i>HBA1</i> locus.....	48
Figure 3.9	VEZF1 bigwig tracks (merged) with MACS2 at <i>HBA1</i> locus.....	49
Figure 3.10	VEZF1 bigwig tracks with MACS2 at <i>TAL1</i> locus.....	50
Figure 3.11	Bigwig tracks with MACS2 at <i>TAL1</i> locus.....	52
Figure 3.12	Comparison between CUT&Tag data and CHIP-Seq data.....	53
Figure 3.13	Association between VEZF1 peaks, GATA2 peaks, promoters, and enhancers.....	54
Figure 3.14	Positional correlation graphs of histone modification enrichment at VEZF1 peaks.....	56

List of tables

Table 2.1	Preparation of stock solutions.....	26
Table 2.2	Preparation of nuclear protein extract dilutions.....	28
Table 2.3	Preparation of the samples.....	28
Table 2.4	QC and sequencing output for the CUT&Tag samples.....	31
Table 2.5	Mapping states of Bowtie2.....	32
Table 2.6	Information on the antibodies and their dilutions or volumes in different experiments.....	34
Table 3.1	Detecting the concentration of new anti-VEZF1 IgG.....	38
Table 3.2	Duplicate reads in the CUT&Tag data sets.....	47
Table 3.3	VEZF1-like motifs enriched at CUT&Tag peaks.....	58
Table 3.4	GATA motifs enriched at CUT&Tag peaks.....	59
Table 3.5	Summary of motifs enriched at VEZF1 and GATA2 binding sites.....	60
Table 3.6	CCAAT box-like motifs enriched at CUT&Tag peaks.....	61

Acknowledgments

Firstly I would like to express my gratitude to the University of Glasgow for providing me an opportunity to study here.

I would like to thank my supervisor Dr Adam West for his guidance and support on the whole project. I would also like to thank Dr Katherine West for the advice and assistance on my experiments and thesis, and teaching me the process of bioinformatic analysis. They really put a lot of effort into my project.

I'm grateful to Dr John Pediani for providing me an opportunity and teaching me how to use the EVOS system, so that I could get nice images of fluorescence.

I want to thank Dr Aqeel Taqi. He helped me get familiar with the institute, the lab, and even the life in Glasgow. I asked him many times for help with my questions and difficulties because he is the only person who work with me in the lab. He was always willing to help me even when he was busy with his work.

Finally, I want to thank my parents for their material and spiritual support during the last four years.

Author's Declaration

The research reported within this thesis is my own work, except where otherwise stated, and has not been submitted for any other degree.

Erxuan Wang

Abbreviation

AML	Acute myeloid leukemia
AP-1	Activator protein 1
bHLH	Basic helix-loop-helix
BSA	Bovine serum albumin
BVEC	Blood vascular endothelial cell
bZIP	Basic leucine zipper
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CTCF	CCCTC-binding factor
CUT&RUN	Cleavage under targets and release using nuclease
CUT&Tag	Cleavage under targets and tagmentation
DBD	DNA-binding domain
DMEM	Dulbecco's Modified Eagle Medium
DTT	Dithiothreitol
EC	Endothelial cell
EDTA	Ethylenediaminetetraacetic acid
EHT	Endothelial-to-hematopoietic transition
EMSA	Electrophoretic mobility shift assay
ESC	Embryonic stem cell
ETS	E26 transformation specific
FDR	False discovery rate
FOX	Forkhead box
HBA1	Hemoglobin subunit alpha 1
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
hET-1	Human endothelin-1
HMM	Hidden Markov model
HMVEC	Human dermal microvascular endothelial cell
HSC	Hematopoietic stem cell
HSPC	Hematopoietic stem and progenitor cell
H3K27ac	Histone H3 lysine 27 acetylation
H3K4me1	Histone H3 lysine 4 monomethylation
H3K4me3	Histone H3 lysine 4 trimethylation
IAHC	Intra-aortic haematopoietic clusters
IF	Immunofluorescence
KCl	Potassium chloride
KLF	Krüppel-like factor
LDS	Lithium dodecyl sulfate

LEC	Lymphatic endothelial cell
LINoCR	LPS-Inducible Non-Coding RNA
LVEC	Lymphatic vascular endothelial cell
MAZ	Myc-associated zinc finger
MDS	Myelodysplastic syndrome
MgCl ₂	Magnesium chloride
MOPS	3-(N-Morpholino)propanesulfonic acid
NaCl	Sodium chloride
NFDM	Non fat dried milk
NF-Y	Nuclear factor Y
NGS	Next generation sequencing
NHR	Nuclear hormone receptor
NK	Natural killer
OP18	Oncoprotein18
pA	Protein A
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PCV	Packed cell volume
PFA	Paraformaldehyde
PVDF	Polyvinylidene fluoride
QC	Quality control
RBR _i	RNA-interaction
RPKM	Reads per kilobase per million
RPMI	Roswell Park Memorial Institute
SCL	Stem cell leukemia
SDS	Sodium dodecyl sulfate
SE	Super-enhancer
SP1	Specificity protein 1
SPRI	Solid-phase reversible immobilization
TAD	Topologically associating domain
TAL1	T-cell acute leukemia protein 1
TEC	TAL1-erythroid complex
TF	Transcription factor
TFBS	Transcription factor binding site
TSH	Thyrotropin
TSS	Transcription start site
VEGF	Vascular endothelial growth factor
VEZF1	Vascular endothelial zinc finger 1

ZIC	Zinc finger protein of the cerebellum
ZF	Zinc finger

Chapter 1 Introduction

1.1 Development of the vascular system

The vasculature is a complex network consisting of arterial, venous, and lymphatic vessels, which provide delivery of nutrients and metabolites for all organs. The formation of the vascular system begins in embryonic development, which is the first identifiable structure of developing mammalian embryos. Two distinct mechanisms of blood vessel formation in the early embryo have been described. Vasculogenesis refers to the differentiation from angioblasts to endothelial cells followed by formation of heart and the first primitive vascular plexus. Angiogenesis is the remodeling and expansion of the vascular network from preexisting endothelial cells (Choi, 2002; Patan, 2004).

Vasculogenesis first occurs in the embryonic yolk sac of mammalian embryos and continues throughout development of embryo proper. During gastrulation, embryonic ectodermal (epiblast) cells are recruited to the primitive streak, a transient embryonic structure located on the posterior side of the embryo. At the primitive streak, a transition from epiblast cells to mesenchymal cells occurs. The mesenchymal cells subsequently migrate and colonize between the visceral endoderm and epiblast, thereby forming either mesoderm or definitive endoderm (Tam and Behringer, 1997; Goldie et al., 2008). In the yolk sac, visceral endoderm releases soluble signals which targets the underlying mesoderm and induces the formation of primitive endothelial and hematopoietic cells, the first differentiated cell types to be produced in the mammalian embryo (Goldie et al., 2008).

The mesoderm cells in the yolk sac wall coalesce to form blood islands, which subsequently occurs in body stalk and chorion. Then the cells around the blood island differentiate into endothelial cells while the cells in center of the blood island differentiate into hematopoietic stem cells (HSCs). The tubules extend and fuse to form a primitive network known as an extraembryonic capillary or vascular plexus. In addition, the mesenchymal cells surrounding the embryonic mesenchymal cleft also differentiate into endothelial cells (ECs) and form intraembryonic vascular plexus. The extraembryonic and intraembryonic vascular plexuses connect through body stalk, which enables HSCs to enter the embryo. Later the mesenchymal cells surrounding the vascular differentiate into smooth muscle cells and pericytes and form intact structure of artery and vein (Figure 1.1).

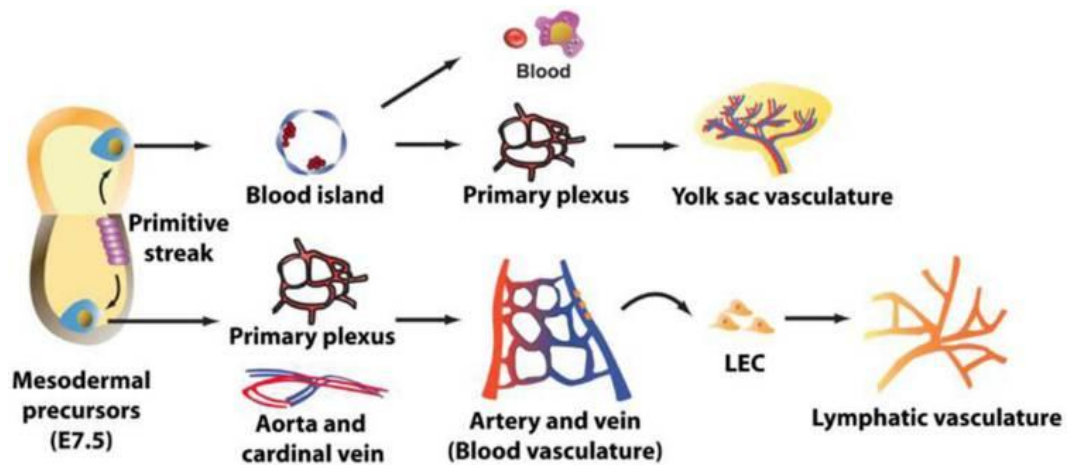


Figure 1.1 Stepwise development of vessels of the three circulations.

Development of vessels start with mesodermal precursor cells, which form respectively blood Island in extraembryonic yolk sac or primary plexus, aorta and cardinal vein in embryo proper. Primitive blood cells and endothelial cells (ECs) are differentiated in blood island. ECs form the vascular primary plexus, which is then remodeled into the yolk sac vasculature. While the primary plexus, aorta and cardinal vein further form arteries and veins. Some ECs are transformed into lymphatic endothelial cells (LECs) by cell fate decision and develop into lymphatic vessels. Taken from Park et al., 2013.

1.2 VEZF1

Our lab has a long standing interest in the DNA binding factor Vascular Endothelial Zinc Finger 1 (VEZF1) due to its roles at the chicken beta globin insulator and in the regulation of hematopoietic gene expression. However, recent work in our lab has demonstrated that it also plays a key role in the control of vasculogenesis (Rivera Gonzalez, 2017).

VEZF1 encodes a 518 amino acid nuclear protein that contains six zinc finger motifs of the Cys2/His2 (Krüppel-like)-type. It is expressed across a broad range of somatic tissues and cell lines including chicken erythrocytes (Koyano-Nakagawa et al., 1994; Ruslan S. Strogantsev, 2009; Lewis et al., 1988). Chromatin immunoprecipitation assays have shown that it binds to the chicken beta globin HS4 insulator, and chromatin barrier assays showed that it is essential for the ability of HS4 to act as a barrier to the spread of heterochromatin (Recillas-Targa et al., 2002). VEZF1 acts by inhibiting the spread of DNA methylation (Dickson et al., 2010), and works alongside USF, which recruits active histone modification complexes to act as a road block to heterochromatin. The barrier activity of VEZF1 and USF is complemented by the enhancer blocking activity of CTCF, and these factors combine to give the powerful insulator function of HS4.

Chromatin immunoprecipitation, RNA interference and CRISPR knockout of VEZF1 in K562 cells have revealed that VEZF1 binds to promoters and enhancers of a large number of genes with diverse functions, and contributes to erythroid gene regulation (Ruslan S. Strogantsev, 2009; Low, 2013; Al-Hosni, 2016). However, it is not essential for the full erythroid gene regulatory programme (Al-Hosni, 2016; Rivera Gonzalez, 2017). To investigate this further, the role of VEZF1 was studied during the erythroid differentiation of human embryonic stem cells (hESCs) *in vitro* (Rivera Gonzalez, 2017). Analysis of VEZF1-knockout hESCs revealed that VEZF1 was not required for human erythropoiesis, and its expression actually inhibited commitment down the erythroid lineage. In contrast, VEZF1 was required for efficient differentiation of endothelial cells in this system (Rivera Gonzalez, 2017). It was therefore hypothesised that VEZF1 is a key regulator of the vascular endothelial gene regulatory programme.

A role of VEZF1 in vasculogenesis is consistent with studies showing that it is upregulated in vascular endothelial cells during vasculogenesis in the developing embryo (Xiong et al., 1999; Kuhnert et al., 2005; Zou et al., 2010). Furthermore, several studies have implicated VEZF1 in function. For example, VEZF1 promotes the formation of the vascular network by binding at the promoter associated CpGi of the antiangiogenic factor *Cited2* gene and repressing the expression of *Cited2* in endothelial cells (AlAbdi et al., 2018). VEZF1 is also associated with the maintenance of vascular integrity, which means the junctions between vascular endothelial cells are compromised in its absence. Loss of a single VEZF1 allele results in an incompletely penetrant phenotype characterized by lymphatic hypervascularization and haemorrhaging and edema in the jugular region (Kuhnert et al., 2005). Furthermore, VEZF1 physically interacts with the GTP-bound form of RhoB, and this complex drives transcription of target set of genes in blood vascular endothelial cells (BVECs) and lymphatic vascular endothelial cells (LVECs), thereby regulating distinct blood and lymphatic endothelial responses to different pathological stimuli (Gerald et al., 2013).

VEZF1 protein expression is believed to be regulated by specific MicroRNAs (miRNAs/miRs), a class of small (≈ 22 nucleotides) non-coding single-stranded RNAs that are necessary for postnatal angiogenesis (Suárez et al., 2008). MiR-191 targets VEZF1 and then affects the expression of downstream genes by regulating their mRNA levels. In patients with acute ischemic stroke and damaged endothelial cells,

MiR-191 has been shown to inhibit angiogenesis by targeting VEZF1 (Du et al., 2019). VEZF1 is also directly suppressed by miR-382-5p. In osteosarcoma, VEZF1 suppression by miR-382-5p can inhibit the development and progression of the tumour (Wu et al., 2021).

1.3 Chromatin binding by VEZF1

ChIP-seq in K562 cells revealed that VEZF1 binds to a major proportion of gene promoters that are active, or primed to be active (Low, 2013). These include cell type-specific and housekeeping genes. Intriguingly, VEZF1 is always located in the nucleosome depleted regions of these promoters and overlaps RNA polymerase II binding immediately upstream of the transcription start site (TSS). VEZF1 does not have any classical transcription activation function, so it was hypothesized that VEZF1 plays a role in establishing the specific chromatin structures observed at transcriptionally active gene promoters. Furthermore, VEZF1 interacts with promoters that have similar G-rich SP1-like motifs to the HS4 insulator, thereby protecting genes with diverse functions from methylation (Dickson et al., 2010).

VEZF1 was also shown to bind to a range of enhancers in a tissue-specific manner (Low, 2013; Ruslan S. Strogantsev, 2009). In K562 cells, comparison with ENCODE data revealed VEZF1 enhancer binding was associated with a complex of erythroid-specific factors that regulate expression of a subset of erythroid genes: *GATA1*, *TAL1*, *E2A*, *LMO2* and *LDB1*.

Electrophoretic mobility shift assays (EMSA) revealed that VEZF1 binds most strongly to homopolymeric runs of G bases (Low, 2013). This is supported by motif analysis of ChIP-seq data, as when binding sites were stratified by VEZF1 signal strength, the enriched motif in the top 5% of peaks was a run of 9 (dG.dC). This motif was most strongly associated with promoter-bound VEZF1 peaks. Weaker VEZF1 peaks sites tended to be enriched in divergent GGGGNGGGG sites, and these were most strongly associated with enhancer-bound VEZF1 peaks. EMSA assays demonstrated that VEZF1 can bind to GGGGAGGGG sites *in vitro*, but does not bind GGGGTGGGG or GGGGCGGGG sites efficiently (Low, 2013).

Based on this data, it was hypothesised that VEZF1-co-operates with erythroid-specific transcription factors to enable binding to weaker sites with degenerate sequences in K562 cells. In particular, the

co-localisation of GATA motifs with VEZF1 sites at the chicken beta globin locus implicates GATA1 as being a key factor in regulating VEZF1 binding (Low, 2013; Ruslan S. Strogantsev, 2009).

In order to investigate the role of VEZF1 in the regulation of endothelial cell differentiation and vascular function, a key step will be to map its binding sites throughout the genome, and to relate these binding sites to those of other factors that may co-operate with VEZF1 to regulate transcription or chromatin organisation.

1.4 Transcription factors regulating endothelial cell differentiation

Several lines of evidence suggest that haematopoietic cells and endothelial cells develop from a common mesodermal progenitor termed the hemangioblast or hemangiogenic endothelium (Figure 1.2). Several transcription factors appear to play key roles in determining endothelial versus haematopoietic differentiation such as SCL/TAL1, and GATA1/2 (Figure 1.2).

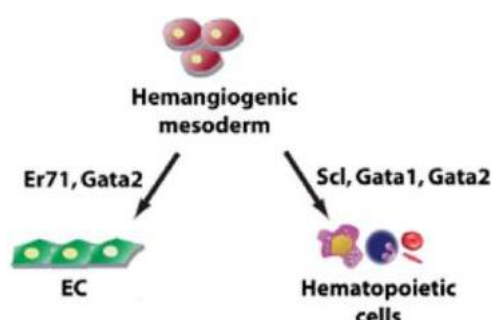


Figure 1.2 Model of hemangiogenic mesoderm cell differentiation by regulation of transcription factors.

SCL/TAL1, and GATA1/2 are essential TFs in the determination of endothelial and hematopoietic lineages. Taken from Park et al., 2013.

1.4.1 GATA2

GATA1 and GATA2 are of particular interest as potential co-factors for VEZF1 in the regulation of haematopoeitic versus endothelial cell differentiation. GATA transcription factors belong to the zinc finger (ZF) superfamily, which has been highly conserved throughout evolution. GATA1 was first identified in erythroid cells as Eryf1 in 1988, with binding sites found within the enhancer region of the chicken f-globin locus and human globin gene (Evans et al., 1988). The sequence of core motifs that Eryf1 recognizes is W(A/T)GATAR(A/G) (Tsai et al., 1989). Subsequently, Eryf1, as the predominant

protein binding to these motifs, was also known as GATA1 (Orkin, 1995). The sequence of an alternative consensus motif that is mainly recognized by GATA2 and GATA3 is AGATCTTA (Ko and Engel, 1993). There are six members of GATA family in vertebrates (GATA1-6), and they are divided into two subgroups, which are GATA1/2/3, and GATA4/5/6 (Whitcomb et al., 2020). GATA1/2/3 are essential for hematopoietic differentiation, including the development of hematopoietic progenitors, erythroid, and T-lymphoid cells (Khandekar et al., 2007).

GATA1 and GATA2 are essential for haematopoietic development, and they regulate the expression and function of multiple transcription factors. GATA2 and GATA1 cooperates with stem cell leukemia (SCL)/T-cell acute leukemia protein 1 (TAL1), a key regulator of hematopoiesis. SCL recognizes E box DNA motif (CANNTG) and forms an obligate heterodimer with ubiquitously expressed class I bHLH E protein. Then two non-DNA-binding proteins, LMO1/2 and LDB1, are recruited and assemble into SCL core complex. LMO2 subsequently recruits additional transcription factors and cofactors, including GATA1/2, thus endow the pentameric complex with multiple functions (Porcher et al., 2017; El Omari et al., 2013). Interestingly, GATA motifs and E boxes usually occur as composite elements (Katsumura and Bresnick, 2017), and SCL preferentially occupies the elements that occupied by GATA2 (Wozniak et al., 2008). Moreover, GATA motifs are the most frequent SCL-bound sequences (Porcher et al., 2017), which indicates a decisive role of GATA2 in SCL occupancy. Apart from SCL, LYL1, another paralogous hematopoietic regulators, is also coregulated by GATA2 and Ets (Chan et al., 2007).

There is evidence that GATA2 plays important roles in both endothelial cell differentiation and haematopoiesis (Park et al., 2013). The expression of GATA2 is detected in embryonic stem (ES) cells, endothelial cells, haematopoietic progenitors, early erythroid cells, mast cells and megakaryocytes. GATA2 drives the expression of the VEGF-A receptor FLK-1/KDR, making the mesodermal cells responsive to vascular endothelial growth factor (VEGF)-A signaling, thereby inducing differentiation into the endothelial cells that form the dorsal aorta. Several studies show that GATA2 regulates the expression of key endothelial genes in human dermal microvascular endothelial cells (HMVECs). GATA2 null mice die by embryonic days 10-11 (E10-E11), which is thought to be because of severe anaemia caused by profoundly defective haematopoietic stem or progenitor cells differentiated from GATA2 null ES cells (Tsai et al., 1994). GATA2 is subsequently involved in the regulating the transcriptional network

that governs the transition from endothelial cells to hematopoietic cells that takes place in the floor of the dorsal aorta (Dobrzycki et al., 2019, Koyunlar et al., 2023).

1.4.2 CTCF

CTCF (CCCTC-binding factor) and VEZF1 are essential components of the the chicken HS4 beta globin insulator (Bell et al., 1999). While VEZF1 works with USF to prevent the spread of heterochromatin, CTCF acts as an enhancer blocker, preventing enhancer sequences from inappropriately activating gene expression (Recillas-Targa et al., 2002; Dehingia et al., 2022). CTCF's enhancer blocking activity stems from its ability to anchor chromatin loops, or topologically associating domains (TADs), through interaction with cohesin. The boundaries of TADs are defined by a pair of convergent CTCF motifs (Dixon et al., 2012; Rao et al., 2014; Wutz et al., 2017). The ring-shaped structure of cohesin consists of three subunits, SMC1, SMC3, and SCC1 (Haering et al., 2008). In the loop extrusion model, cohesin is loaded onto chromatin by the SCC2/SCC4 complex and moves along the chromatin fibre to extrude a loop (Haarhuis et al., 2017; Gassler et al., 2017). The movement stops when cohesin encounters two successive CTCF proteins bound separately to a pair of convergent motifs. During the extrusion, the enhancer and promoter in the same TAD are likely to interact, ensuring the specific activation of gene (Dehingia et al., 2022). PDS5 and WAPL cooperate to release cohesin from chromatin, restricting the extension of chromatin loops (Wutz et al., 2017). The balance between WAPL and SCC2/SCC4 complex maintains the dynamic of TADs, and the correct structure of the chromosome (Haarhuis et al., 2017).

CTCF consists of 11 ZF domains. ZFs four to seven play an essential role in DNA binding, while other ZFs are deployed in different combinations at different sites (Lobanenkov et al., 1990, Renda et al., 2007). CTCF binding to DNA is regulated by multiple factors, including DNA methylation, histone modifications, chromatin openness, histone variants and RNA (Dehingia et al., 2022; Wen et al., 2020). For example, the interaction of RNA and CTCF stabilizes CTCF binding and the formation of chromatin loops (Saldaña-Meyer et al., 2014; Saldaña-Meyer et al., 2019). However, the transcription of long non-coding RNAs such as LPS-Inducible Non-Coding RNA (LINoCR) and Jpx RNA evict CTCF and cohesin from chromatin (Lefevre et al., 2008; Sun et al., 2013). Furthermore, the function of CTCF is regulated by post-translational modification including phosphorylation and poly(ADP-ribosylation) (El-Kady and Klenova, 2005; Yu et al., 2004).

As an essential chromatin regulator, mutations in the CTCF gene and its binding sites lead to loss of the TAD boundary activity, resulting in cancer and various neurological diseases (Dehingia et al., 2022).

1.5 Epigenetics and histone modifications

The term epigenetics refers to changes in gene function that are inherited through mitosis or meiosis, and are independent of changes in the DNA sequence (Wu and Morris, 2001). The first epigenetic mark was identified in 1962, which was histone methylation. More marks were identified subsequently, including DNA methylation, histone acetylation, phosphorylation, ubiquitylation, sumoylation, and ADP ribosylation, etc (Peixoto et al., 2020). As a trending research field in recent years, epigenetics has been proven to regulate micro and macro processes in many aspects, from gene expression to ontogeny. In our research, three histone modifications are focused on, namely histone H3 lysine 4 monomethylation (H3K4me1), histone H3 lysine 4 trimethylation (H3K4me3), and histone H3 lysine 27 acetylation (H3K27ac). These three histone modifications are enriched to different degrees at promoter and enhancer regions, and can be used, along with other chromatin features, to help identify regulatory elements (Ernst and Kellis, 2010; Ernst et al., 2011). Studying the relationship between VEZF1 binding, erythroid transcription factors, and the surrounding epigenetic landscape will be an important aspect of learning how VEZF1 regulates chromatin structure and gene expression.

1.5.1 H3K4me1

H3K4me1 is mainly enriched at the enhancer regions as an activation mark, which results in active or repressive effects depending on interactions with H3K27ac or H3K27me3 (Zhang et al., 2021). The enhancers that only have H3K4me1 are primed enhancers (Sharifi-Zarchi et al., 2017). They are in a pre-activated state and are activated with the acquisition of H3K27ac, the other activation mark. The enhancers where the H3K4me1 interacts with H3K27ac are active enhancers. They are also bound by mediator, a transcriptional coactivator complex, thereby promoting the transcription of target genes (Heinz et al., 2015). Interestingly, mediator forms a complex with cohesin and is involved in the formation of chromatin loops, which implies a potential association between CTCF and H3K4me1 (Kagey et al., 2011). The enhancers with H3K4me1 and the repression mark H3K27me3 are poised enhancers

(Heinz et al., 2015). They convert to active enhancers when H3K27me3 is removed and H3K27ac is acquired (Karnuta and Scacheri, 2018; Maurya, 2021).

1.5.2 H3K4me3

H3K4me3 is a predominant mark of active or poised promoters, whose existence defines the active state of gene (Santos-Rosa et al., 2002; Calo and Wysocka, 2013). However, the removal of H3K4me3 has no instructive effect on the transcription of most genes, which indicates that the association between H3K4me3 and transcription is more complex than activation and being activated. H3K4me3 is mediated by Set1 or Set1-like methyltransferase complexes, and the recruitment of them are transcription-dependent. This suggests that H3K4me3 is more likely to be the result of transcription instead of the cause. In fact, H3K4me3 has been proved to be linked to transcript degradation, transcriptional responsiveness, and transcriptional consistency (Howe et al., 2017). The respective enrichment of H3K4me1 and H3K4me3 at enhancers and promoters is regulated by DNA methylation through a mechanism described as a seesaw. When DNA is hypermethylated, the level of both H3K4me1 and H3K4me3 are low, which marks the inactive genomic regions. When DNA is intermediately methylated, the level of H3K4me1 is high and the level of H3K4me3 is low, which marks the enhancer regions. When DNA is hypomethylated, the level of H3K4me1 is low and the level of H3K4me3 is high, which marks the promoter regions (Sharifi-Zarchi et al., 2017).

1.5.3 H3K27ac

The function of H3K27ac is closely associated with H3K4me1 and H3K4me3. H3K27ac is mainly enriched at the enhancers and promoters and marks active genes (Zhang et al., 2021). The role of H3K27ac as an activation mark in active enhancers has been introduced above. H3K27ac is also a mark of super-enhancers (SEs). SEs are regions containing multiple clustered enhancers in the genome which drive the specific expression of genes associated with determination of cell identity and fate. SEs is regulated through Notch signaling pathway, and H3K27ac is enriched at NOTCH1 sites, which occur abundantly in SEs (Khan and Zhang, 2016).

1.6 CUT&Tag

The most well known method for identifying and mapping the distribution of chromatin features and histone modifications is chromatin immunoprecipitation followed by sequencing (ChIP-seq). In ChIP assays, chromatin is fragmented and immunoprecipitated with antibodies coupled to agarose beads, then the DNA fragments are purified. Typically, samples are then sent to a specialised facility for library preparation and high throughput sequencing (Rodríguez-Ubreva and Ballestar, 2014). However, the application of ChIP-seq is limited due to its requirement for large sample sizes (typically 10^6 - 10^7 cells), high background signal, and chromatin accessibility artefacts (Kaya-Okur et al., 2020). In addition, the costs of library preparation and high throughput sequencing (typically £250 per sample) limit its application for large studies with multiple samples and replicates. An alternative approach is to use protein A-enzyme fusion proteins, which are recruited to genomic target sites by antibodies that recognise the protein or histone modification being studied. Cleavage under targets and release using nuclease (CUT&RUN) was established based on this strategy, and uses protein A-MNase fusion proteins to induce DNA cleavage at antibody target sites (Kaya-Okur et al., 2020). Although CUT&RUN greatly reduces the number of cells required (100-1000 cells) and lowers the background levels and the cost, it still requires extra steps including polishing of DNA ends and adapter ligation, which take more time and introduce opportunities for variability (Kaya-Okur et al., 2019). Cleavage under targets and tagmentation (CUT&Tag) is a novel method that overcomes the limitations of ChIP-seq and CUT&RUN (Hatice S Kaya-Okur et al 2019). CUT&Tag introduces protein A-Tn5 transposase fusion proteins, which recognize the antibody binding to the target factor, cut the genome, and integrate adapters at both ends of the fragments. The DNA fragments undergo purification, amplification via PCR, another purification, and are then ready for sequencing without any additional library preparation steps (Figure 1.3). CUT&Tag has higher sensitivity than CUT&RUN due to the high efficiency of tag integration at the target sites. Compared to ChIP-seq, CUT&Tag requires fewer cells, so is suitable for single-cell platforms, and takes less time, as the whole procedure can generate a sample ready for sequencing within one day. The high signal-to-noise ratio means that only 3 million reads are required to map histone modifications across a mammalian genome compared to a minimum of 30 million typically used for ChIP-seq. Sequencing costs of only £45 per sample, in our experience, make this method much more suitable for large scale studies requiring multiple samples. Due to the rapid procedure and low cost, CUT&Tag has the potential to

overtake ChIP-seq as the most widely used method for mapping the genomic profiles of DNA binding proteins and histone modifications (Kaya-Okur et al., 2019).

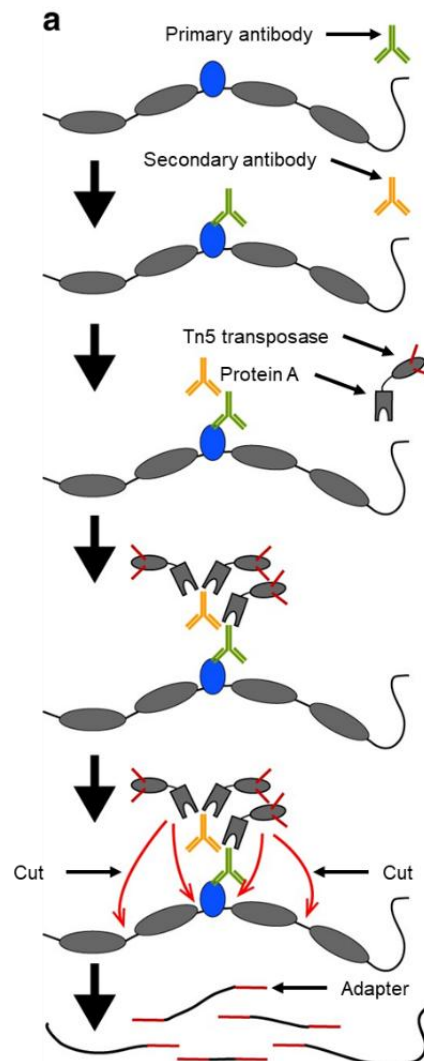


Figure 1.3 In situ tethering for CUT&Tag chromatin profiling.

The steps in CUT&Tag. Primary antibody (green) binds to the target chromatin protein (blue). Secondary antibody (orange) binds to the primary antibody and provide more binding sites for pA-Tn5 transposome (gray boxes), which cut genome at both sides of the target protein and integrates adapters (red) at chromatin protein binding sites after activated by addition of Mg^{2+} . Taken from Kaya-Okur et al., 2019.

1.7 Aims of the project

The first aim of this project is to test and validate the use of CUT&Tag technology as an alternative to ChIP-seq for the study of sequence-specific DNA binding proteins such as VEZF1. The ease of use and low cost of CUT&Tag could enable extensive analysis of the role that VEZF1 plays in regulating the differentiation of endothelial and hematopoietic cells. The co-operation and competition of VEZF1 with other DNA binding proteins and epigenetic modifiers has not been explored, and is likely to be fundamental to understanding the regulatory network that controls vascular differentiation. This would require extensive profiling of DNA binding factors and chromatin marks in a series of differentiating cells, so it is important to test the CUT&Tag technology prior to embarking on such a large scale project.

The second aim is to test the hypothesis that co-operation of VEZF1 with erythroid transcription factors such as GATA1/2 enables it to bind to degenerate sequence motifs *in vivo* that it is unable to bind *in vitro*. Previous work has indicated that VEZF1 prefers to bind to runs of homopolymeric G, and that these motifs can be identified at the most highly enriched VEZF1 peaks at promoters. However, weaker VEZF1 sites, and those at enhancers, follow a GGGGNGGGG pattern that is weakly bound *in vitro*, indicating that other factors may promote VEZF1 binding at these locations. By identifying VEZF1 sites that also bind GATA2, we aim to test the hypothesis that GATA2 modulates the DNA binding specificity of VEZF1 *in vivo*.

1.8 Objectives

1) Test and optimise antibodies for CUT&Tag.

It is important to check that antibodies used for CUT&Tag are specific for the intended protein. Furthermore, it has been reported that immunofluorescence assays are the most suitable method for optimising protein concentration prior to using them in the CUT&Tag procedure. Therefore, antibodies to VEZF1, GATA2, CTCF and the histone modifications H3K4m3, H3K4me1 and H3K27ac will be tested and optimised by western blotting and immunofluorescence.

2)How does data generated by CUT&Tag compare to ChIP-seq data in K562 cells?

The quality of the data generated by CUT&Tag will be compared to that generated by ChIP-seq through consideration of peak location, peak height and background signal. CUT&Tag data for VEZF1 in K562 cells will be compared to ChIP-seq data previously generated by the lab. CUT&Tag histone modification profiles will be compared to those previously generated by the lab, and also compared with publicly available data from K562 cells generated as part of the ENCODE project. The data will be compared by visually examining tracks displayed on the UCSC genome browser.

3)How does the binding of VEZF1 compare to that of GATA2 and CTCF in K562 cells, as assayed by CUT&Tag?

Previous work has shown that VEZF1 binds at active promoters and enhancers in K562 cells, and comparison with ENCODE data indicated that a subset of these VEZF-bound enhancers are also bound by a complex of erythroid-specific transcription factors. By performing CUT&Tag for VEZF1 and GATA2 in the same batch of K562 cells, we can address this question more rigorously, and investigate the proportion of VEZF1 sites that are also bound by GATA2. Similarly, the locations of VEZF1 peaks will be compared with those of CTCF to identify any putative insulator regions that could have both enhancer blocking and barrier activity.

4)Does VEZF1 bind to different consensus DNA sequence motifs at promoters and enhancers, and does the presence of GATA 2 alter the consensus motifs?

VEZF1 peaks will be stratified by whether they overlap with promoters or enhancers, and whether they overlap with GATA2 peaks. Motif finding will then be used to identify enriched DNA sequences at these sites to investigate whether the presence or absence of GATA2 modulates the DNA binding activity of VEZF1.

Chapter 2 Materials and Methods

2.1 Cell culture

2.1.1 Cell thawing

The cryovial with cells was taken out from liquid N₂ tank and incubated at 37°C until only a small ice pellet remained. The vial was sprayed down with ethanol, wiped, and placed into the hood. The cell suspension in the vial (~ 1 ml) was transferred into T25 flask. Cold fresh medium (4 ml) was slowly added into the flask, at a rate of about 1 drop every 10 seconds, swirling occasionally. Another 5 ml of cold fresh medium was added. The flask was placed in incubator at 37°C. Cells were spun down after 6-12 hours and resuspend in fresh, prewarmed medium in a new T25 flask.

2.1.2 Culture conditions

The K562 cells were grown in T25 flasks with 7 ml of RPMI 1640 Medium (Gibco™, 61870036, 10% FBS, 1% p/s). The cell incubator was set to 37°C, 5% CO₂.

The HEK293 cells were grown in T25 flasks with 7 ml of Dulbecco's Modified Eagle Medium (DMEM, Gibco™, 10% FBS, 1% p/s). The cell incubator was set to 37°C, 5% CO₂.

2.1.3 Cell counting

Cell suspension (40 µl) was transferred into a 1.5 ml Eppendorf tube followed by addition of 160 µl Trypan blue dye. The coverslip was placed on the hemocytometer and 10 µl of the mixture was added at both edges of the coverslip to make it fill the counting areas. The two counting areas were observed under the microscope and the cells in grids located on the four corners of each counting area were counted. Cells per ml = ((cells in the first counting area + cells in the second counting area) / (number of counting areas)) x 10⁴ x dilution factor (4).

2.1.4 Cell passaging

Cells were passaged every three days. Cells were seeded into a new container at 1.5×10^5 cell/ml (typically 1:5). The fresh medium and cell suspension were added into new flasks (volumes used for passage were determined according to the confluency of the cells) and mixed by gentle rocking. The flasks were incubated at 37°C.

2.2 Nuclear extract preparation

2.2.1 Preparation of stock solutions and buffer solutions

Buffer A and Buffer C were prepared as the following tables (Table 2.1) ahead of time and pre-chilled on ice before use. One Protease inhibitors tablet (Roche, 5892791001) was resuspended in 5 ml dH₂O and used as a 2X stock. The stock was added into the buffer just prior to use.

Buffer A	Final concentration	Buffer C	Final concentration
HEPES (pH 7.9)	10 mM	HEPES (pH 7.9)	20 mM
MgCl ₂	1.5 mM	glycerol	25%
KCl	10 mM	NaCl	0.42 M
DTT	0.5 mM	MgCl ₂	1.5 mM
		EDTA	0.2 mM
		DTT	0.5 mM

Table 2.1 Preparation of stock solutions.

2.2.2 Harvesting cells

Around 20 ml of cell suspensions were used to prepare nuclear extract each time. Cells were grown for two days before harvesting to ensure they were confluent. Cell suspensions were pelleted at 450 xg for 5 minutes. The pellet was resuspended by gentle pipetting in the same volume as the original culture of 1X PBS. The cells were pelleted at 450 xg for 5 minutes and the supernatant was discarded. The residual supernatant was removed by pulse spin and the packed cell volume (PCV) was estimated.

2.2.3 Swelling the cells

The pellet was resuspended by gentle pipetting in 20 PCV of ice cold 0.67X PBS. The cell suspension was incubated on ice for 10 minutes. The cells were counted at this point.

2.2.4 Cell disruption

The cells were pelleted at 450 xg for 5 minutes and the supernatant was discarded. The residual supernatant was removed by pulse spin. The pellet was resuspended by gentle pipetting in at least 5 PCV of ice cold buffer A. The cells were lysed by using five gentle passes through a narrow gauge needle (G25 or G27) using a syringe. The lysis was checked under a microscope using trypan blue, which should be 80-90%. The nuclei were pelleted for 20 minutes at 11000 xg at 4°C and the supernatant was transferred into a fresh tube. The residual supernatant was removed by pulse spin.

2.2.5 Salt extraction of nuclear proteins from chromatin

The crude nuclei pellet was resuspended in 2/3 PCV of Buffer C. The nuclei were disrupted by using ten gentle passes through a narrow gauge needle (G25 or G27) using a syringe and agitated gently for 30 minutes at 4°C. The nuclear debris was pelleted for 10 minutes at 16,500 xg and the soluble nuclear protein supernatant was transferred into a clean, chilled tube. Then the supernatant was snap-frozen in small aliquots with dry ice or liquid nitrogen and stored at -70°C.

2.3 Bradford Assay

Bovine serum albumin (BSA) solution (5 µl, 20 mg/ml, NEB, B9000S) was diluted with 995 µl of distilled water to 100 µg/ml. Five BSA dilutions with concentrations of 8, 12, 20, 40, and 80 µg/ml were prepared by mixing different volume of the 100 µg/ml BSA and distilled water.

3 sets of nuclear protein extract dilutions were prepared in parallel by mixing different volume of each nuclear protein extract and distilled water in the table below (Table 2.2).

Dilution factor	Extract (μ l)	Distilled water (μ l)	Final volume (μ l)
500	1	499	500
125	4	496	500

Table 2.2 Preparation of nuclear protein extract dilutions.

160 μ l of each BSA standard and the diluted nuclear protein extract was transferred to a 96-well plate. Each mixture was added in triplicate. 40 μ l of the Dye Reagent Concentrate (5X, BIO-RAD, # 500-0006) was added respectively to each well and mixed using a multi-channel pipet. The 96-well plate was incubated at room temperature for 23 min and the absorbance of each well was measured using the TECAN micro plate reader setting a wavelength of 595 nm and 450 nm.

2.4 Western blotting

Western blotting was applied to determine the activity and specificity of antibodies.

The samples were prepared by diluting nuclear extracts respectively with different volumes of dH₂O, followed by addition of LDS Sample Buffer (and Sample Reducing Agent) according to the table (Table 2.3). The mixtures were heated at 70°C in dry bath for 10 minutes.

Nuclear extract or marker	Protein load (μ g)	Volume of nuclear extract (μ l)	dH ₂ O (μ l)	4X NuPAGE™ LDS Sample Buffer (Invitrogen, NP0007) (μ l)	NuPAGE™ Sample Reducing Agent (10X, Invitrogen, NP0004) (μ l)	Total volume (μ l)
K562 Parental (old)	17.82	8.59	3.41	4		16
K562 Parental (new)	7.82	3	7.4	4	1.6	16
K562 KO	17.82	10	2	4		16

Table 2.3 Preparation of the samples.

50 ml of 20X MOPS SDS Running Buffer (Novex, B0001) is diluted with 950 ml of distilled water to prepare 1X SDS Running Buffer.

The NuPAGE™ 4-12% Bis-Tris Mini Gel (Invitrogen, NP0321BOX) were placed in the mini gel tank. 1X MOPS SDS Running Buffer (Novex, B0001) with or without NuPAGE™ Antioxidant (NP0005) was added to the XCell SureLock™ Mini-Cell. Appropriate volume of samples and markers were loaded in the appropriate wells. The gel was run at 200 V constant for 50 minutes .

Once the running was finished, the gel was removed and assembled in a TransBlot cassette for transfer to PVDF membrane. Transfer was carried out using the TransBlot apparatus at 2.5 A for 7 minutes,

according to the manufacturer's instructions (Bio-Rad)

After the transfer program was complete, the membrane was blocked in 5% non fat dried milk (NFDM, "Marvel")/ 1X PBS-T for 30 minutes at room temperature and then at 4°C overnight.

The membrane was rinsed briefly in 1% NFDM/1X PBS-T. Then the membrane was transferred into clean plastic pockets and incubated in primary antibody solution for 1 hour at room temperature.

The membrane were washed 3 times with 1% NFDM/1X PBS-T for 15 minutes each at room temperature and incubated in secondary antibody solution for 1 hour at room temperature. After another 3 washes of 1% NFDM/1X PBS-T and 1 wash of 1X PBS-T, the membrane was placed on a level surface with all the milk removed. The imaging reagent (A38554, SuperSignal, Thermo Scientific™) was prepared and added onto the membrane and incubated for 5 minutes at room temperature. The excess reagent was subsequently removed and the membrane was placed into a clean plastic pocket and covered. Eventually the membrane was imaged on a compatible digital imaging system (G-box imager, Syngene). Following imaging, the pocket was sealed and stored in the fridge.

2.5 Immunofluorescence

HEK cells were grown in the 24-well plate at 37°C and 5% CO₂ overnight. When the cells reached appropriate confluency (around 10⁵ cells per well), the media was discarded and 4% paraformaldehyde (PFA) solution made in PBS was added to fix the cells. The cells were then permeabilised with 0.1% triton-PBS for 10 minutes, blocked with IF buffer (5% horse serum and 0.5% triton-PBS) for 45 minutes, and incubate in primary antibody dilutions at 4°C overnight, and secondary antibody dilutions at room temperature for 2 hours sequentially. Eventually the cells were stained with Hoechst 33342 working solution (1 µg/ml) for 10 minutes and 1X Phalloidin conjugate working solution (prepared by diluting 1000X Phalloidin conjugate Stock solution 1000 times in 1 ml of PBS + 1% BSA) for 1 hour, and imaged with EVOS system (Thermo Scientific™).

2.6 CUT&Tag

2.6.1 Sample preparation

The CUT&Tag procedure was carried out according to the manufacturer's instructions (Active Motif). Briefly, cells were seeded into a new T25 flask at 1.5×10^5 cell/ml and grown under standard conditions (37°C, 5% CO₂) for around 45 hours. The cells were counted and harvested into a 1.5 ml microcentrifuge tube (5×10^5 cells per sample), washed with 1X Wash Buffer, and placed on ice. The cells were then immobilised on Concanavalin A-coated magnetic beads and the cell membrane permeabilised with digitonin. Then the cells were incubated sequentially in primary antibody solutions [contain 1 µg of antibody (0.5 µg and 0.1 µg were also applied on CTCF antibody) diluted in Antibody Buffer, 4°C, overnight], secondary antibody solution [Guinea Pig Anti-Rabbit Antibody (100 µl for each reaction) diluted in Dig-Wash Buffer, room temperature, 60 minutes], and CUT&Tag-IT™ Assembled pA-Tn5 Transposomes (diluted in Dig-300 Buffer, room temperature, 60 minutes). In this step, the Tn5 transposase is recruited to the antibody-bound genome locations via protein A (pA). The addition of Tagmentation Buffer, which contains Mg²⁺, enables the Tn5 transposome enzyme to cut genome and tag the 5' DNA ends with with sequencing adapters. This step was carried out at 37°C for 60 minutes. Following tagmentation, DNA was extracted and purified with DNA Purification Column. PCR was used to amplify DNA fragments. Each sample was amplified using one of four i7 indexed primers as the forward primer, and one of four i5 indexed primers as the reverse primer. Each sample therefore has a unique barcode, allowing up to 16 samples to be multiplexed in the same high throughput sequencing reaction. SPRI Bead clean-up was performed to further purify the PCR products. Samples were quantified by Qubit (Glasgow Polyomics) and were estimated to contain between 7.5 ng and 38 ng of DNA.

2.6.2 Sample sequencing

Samples were sent to Novogene Europe for quality control (QC) analysis and sequencing (table 2.4). Novogene reported that the concentration of DNA as determined by QPCR ranged from 0.7 nM to 13.91 nM. Bioanalyser traces revealed multiple peaks between 200 bp and 600 bp for most samples. Histone modification samples had higher DNA concentrations and stronger bioanalyser peaks than samples for VEZF1, GATA2 and CTCF. Although not all samples passed the Novogene QC filters, none were discarded

at this stage. This is because the DNA samples generated by the CUT&Tag procedure are expected to have a lower concentration and less background signal compared to standard ChIP-Seq samples, so standard QC requirements are not applicable. Sequencing was carried out on all samples using paired-end 150 bp NovaSeq technology, with 4 Gb raw data output requested. The number of reads obtained ranged between 5.9 million and 38 million, and there was no correlation between the number of reads or the read quality and the original DNA concentration or bioanalyser peak QC (Table 2.4).

Sample name	Antibody	antibody amount	rep. number	QPCR (nM)	bioanalyser peak description	Library QC	Raw reads	Raw data (Gb)	base error rate (%)	bases with phred value > 20(%)	bases with phred value > 30(%)
CnT_VEZF1_AGW_1	VEZF1	1 µg	1	0.96	no visible peak	Fail	35,930,348	5.4	0.03	97.21	93.58
CnT_VEZF1_AGW_2	VEZF1	1 µg	2	4.37	small fragment	Fail	32,014,290	4.8	0.03	95.36	91.33
VEZF1_3_1	VEZF1	1 µg	3	0.80	pass	Fail	34,675,678	5.2	0.03	95.57	90.2
VEZF1_3_2	VEZF1	1 µg	4	1.39	pass	Fail	37,445,428	5.6	0.03	95.19	89.39
CtCF_01_1	CtCF	0.1 µg	1	1.47	pass	Fail	24,684,422	3.7	0.03	94.95	87.68
CtCF_01_2	CtCF	0.1 µg	2	0.70	pass	Fail	23,894,938	3.6	0.03	96.85	92.65
CtCF_05_1	CtCF	0.5 µg	1	6.00	multipeak, small fragment	Fail	29,558,128	4.4	0.03	95.67	91.46
CtCF_05_2	CtCF	0.5 µg	2	2.72	pass	Pass	26,574,278	4	0.03	96.59	92.01
CnT_CtCF_2	CtCF	1 µg	1	6.58	multipeak, small fragment	Fail	35,253,722	5.3	0.03	97.62	94.26
GATA2_1	GATA2	1 µg	1	1.62	small fragment	Hold	25,125,704	3.8	0.03	97.14	92.67
GATA2_2	GATA2	1 µg	2	2.29	small fragment	Hold	27,454,674	4.1	0.03	96.77	92.2
GATA2_3	GATA2	1 µg	3	0.79	small fragment	Fail	25,518,066	3.8	0.03	96.63	91.86
CnT_H3K27ac_1	H3K27ac	1 µg	1	10.36	multipeak, small fragment	Fail	35,285,006	5.3	0.03	95	89.01
CnT_H3K27ac_2	H3K27ac	1 µg	2	7.94	multipeak, small fragment	Fail	32,911,468	4.9	0.03	97.54	93.71
H3K27ac_3_1	H3K27ac	1 µg	3	6.46	multipeak	Hold	31,667,286	4.8	0.03	97.09	92.53
H3K27ac_3_2	H3K27ac	1 µg	4	10.20	multipeak, small fragment	Fail	28,238,388	4.2	0.03	97.14	92.87
H3K4me1_1	H3K4me1	1 µg	1	13.90	multipeak	Hold	38,123,078	5.7	0.03	97.14	93.07
H3K4me1_2	H3K4me1	1 µg	2	12.30	multipeak	Hold	15,948,866	2.4	0.03	96.54	91.71
H3K4me1_3	H3K4me1	1 µg	3	12.19	multipeak	Hold	17,227,716	2.6	0.03	96.52	91.59
H3K4me3_1	H3K4me3	1 µg	1	13.91	multipeak, small fragment	Fail	5,914,800	0.9	0.04	92.59	84.07
H3K4me3_2	H3K4me3	1 µg	2	13.30	multipeak	Hold	18,403,174	2.8	0.03	96.09	91.06

Table 2.4 QC and sequencing output for the CUT&Tag samples.

The amount of each antibody used is indicated. Novogene provided data on sample quantification by QPCR, analysis by bioanalyser, and whether the sample passes their QC tests. Novogene also provided data on the sequencing data output as indicated in the column headers.

2.6.3 Next Generation Sequencing (NGS) data processing and analysis

Processing and analysis of NGS sequencing data was performed using tools on the European Galaxy server (usegalaxy.eu) unless otherwise stated. FASTQC was used to assess the quality of the raw data (Andrews, 2010), as it displays the Phred score, GC content, adaptor content and other indicators. Due to high adaptor content, the tool cutadapt was used to remove adapters corresponding to the i5 and i7 index primers used during library preparation by PCR. The sequence CTGTCTTATACACATCT was used to identify 3' (end) adapters, with the adapter and subsequent bases being trimmed. FASTQC was repeated to determine whether adapter trimming had improved the quality of the data.

Bowtie2 was used to align the forward and reverse reads to the Hg19 genome (Langmead et al., 2009; Langmead and Salzberg, 2012), creating bam files. The Hg19 genome is not the newest reference genome, but it was still adopted because the ChIP-Seq data previously generated by the AGW lab was

also aligned to the Hg19 genome. Moreover, there was more publicly available histone modification data aligned to Hg19 on the UCSC genome browser that could be used as comparison. The minimum and maximum fragment lengths were set to 10 and 700 bp respectively, and the “very sensitive local” alignment settings were chosen. The mapping states were preliminary judged according to the proportion of concordant alignment (Table 2.5).

Data set	Aligned concordantly 0 times	Aligned concordantly exactly 1 time	Aligned concordantly >1 times
VEZF1_AGW_1	8.88%	48.73%	42.38%
VEZF1_AGW_2	11.38%	45.13%	43.49%
VEZF1_3_1	13.16%	46.53%	40.32%
VEZF1_3_2	11.66%	50.90%	37.44%
GATA2_1	12.87%	35.86%	51.26%
GATA2_2	8.83%	45.06%	46.11%
GATA2_3	9.52%	44.26%	46.22%
CTCF_2	10.11%	46.59%	43.30%
CTCF_01_1	7.20%	62.14%	30.66%
CTCF_01_2	6.85%	61.27%	31.88%
CTCF_05_1	8.59%	54.77%	36.64%
CTCF_05_2	7.93%	56.82%	35.25%
H3K4me3_1	10.16%	65.43%	24.41%
H3K4me3_2	7.07%	70.12%	22.81%
H3K4me1_1	5.50%	60.35%	34.15%
H3K4me1_2	6.40%	59.23%	34.38%
H3K4me1_3	6.92%	58.39%	34.69%
H3K27ac_1	7.90%	63.07%	29.03%
H3K27ac_2	11.21%	45.38%	43.41%
H3K27ac_3_1	9.71%	49.82%	40.47%
H3K27ac_3_2	10.80%	48.99%	40.21%

Table 2.5 Mapping states of Bowtie2.

In most of the data sets, the proportion of concordant alignment is over 90%, which is satisfying. Some reads were aligned concordantly more than once. A potential reason was that the binding and cutting pattern of Tn5 transposome enzyme might result in fragments with similar sequences. However, these proportions are only for reference. The quality of alignment would be determined based on checking visually on the UCSC genome browser instead of the proportion of concordant alignment.

To display the NGS data on the UCSC genome browser, the tool bamcoverage was used (Ramírez et al., 2016), with a bin size of 10 bp and normalisation to reads per kilobase per million (RPKM). Files were hosted on the Cyverse discovery environment web server for rapid access by the UCSC genome browser. Each bigwig file was checked visually and one sample with low enrichment and high background was discarded: CnT_H3K27ac_2. Bam files from replicates were combined using Samtools merge to give combined bam files for VEZF1, GATA2, CTCF, H3K27ac, H3K4me1 and H3K4Me3.

Peaks of enrichment were identified using the tool MACS2 callpeak with a minimum FDR of 0.05 or 0.01. The narrow peak bed file outputs were uploaded to the UCSC genome browser for visualisation.

Active promoters and enhancers in K562 cells were defined using the data from the Broad Institute ChromHmm analysis of K562 cells (Ernst and Kellis, 2017), which uses histone modification data to predict the functionality of genomic regions. The ChromHmm data set was downloaded from the UCSC genome browser using the table browser functionality and imported into galaxy. Columns were rearranged into standard bed format using the cut tool, and the filter tool was used to filter chromatin segments by category. Segments designated "1_active promoter" and "2_weak_promoter" were combined to give a list of active promoters in K562 cells. Segments designated "4_Strong_Enhancer", "5_Strong_Enhancer", "6_Weak_Enhancer" and "7_Weak_Enhancer" were combined to give a list of enhancers in K562 cells.

Bedtools Intersect intervals was used to obtain lists of overlapping or non-overlapping peaks when comparing two bed files, using a minimum of 1 bp for the overlapping region. To obtain a list of VEZF1 or GATA2 peaks that overlap with enhancers, peaks that overlap with active promoters were identified and removed from the list initially, and the remaining peaks were then overlapped with enhancers. This is because many of the promoters have enhancer-like segments adjacent to them, which are likely to be the shoulders of the promoter histone modification enrichments rather than stand-alone true enhancers.

For sequence motif discovery, 200 bp windows around each peak summit were identified, using the peak summit information in the bed file. The tool bedtools getfasta was used to obtain a list of all the DNA sequences in each of these 200 bp windows. Sequence motifs enriched in each data set were identified using the tool MEME-ChIP (Ma et al., 2014). Tomtom was used to identify transcription factors that are known to bind to motifs identified using MEMECHIP (Bailey et al., 2015).

Comparison of peaks with histone modification data was performed using the ChIP-Seq tool on the Swiss Bioinformatics Resource Portal at epd.expasy.org/cgi-bin/chipseq/ (Ambrosini et al., 2016). The programme ChIP-Cor was used to quantify the enrichment of histone modifications around peak summits, using the global normalisation option. The data was then downloaded and plotted in excel. The programme ChIP-Extract was used to generate the heat maps to display the enrichment data.

2.7 Antibodies

Information on all antibodies used in this study is listed in the table below (Table 2.6).

Antibody (name/catalogue number)	Supplier	Final dilution for western blotting	Final dilution for immunofluorescence	Final volume/dilution for CUT&Tag
VEZF1 Antibody (AGW-3642-old)	AGW's lab	1:5,000		
VEZF1 Antibody (AGW-3642-new)	AGW's lab	1:5,000	1:200	3.85 μ l (1 μ g)
VEZF1 Antibody (AGW-3645-new)	AGW's lab	1:5,000		
VEZF1 Antibody (NBP1-84301-25ul)	Novus Biologicals	1:1,000	1:50	
SP1 Antibody (39058)	Active Motif	1:10,000		
Donkey Anti-Rabbit IgG (H+L) Secondary Antibody, HRP (A16023)	Thermo Fisher Scientific	1:100,000		
Goat anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor™ 488 (A-11008)	Thermo Fisher Scientific		1:500	
GATA2 Antibody (PA1-100)	Thermo Fisher Scientific	1:2,000		
GATA2 Antibody (NBP1-82581)	Novus Biologicals	1:5,000		
GATA2 Antibody (ab153820)	Abcam	1:1,000	1:100	2.63 μ l (1 μ g)
CTCF Antibody (ab188408)	Abcam	1:40,000	1:2000	0.33 μ l (0.5 μ g), 0.066 μ l (0.1 μ g)
H3K4me1 Antibody (ab8895)	Abcam			1 μ l (1 μ g)
H3K4me3 Antibody (ab213224)	Abcam			2.04 μ l (1 μ g)
H3K27ac Antibody (ab177178)	Abcam			0.71 μ l (1 μ g)
Guinea Pig Anti-Rabbit Antibody (53160)	Active Motif			1:100

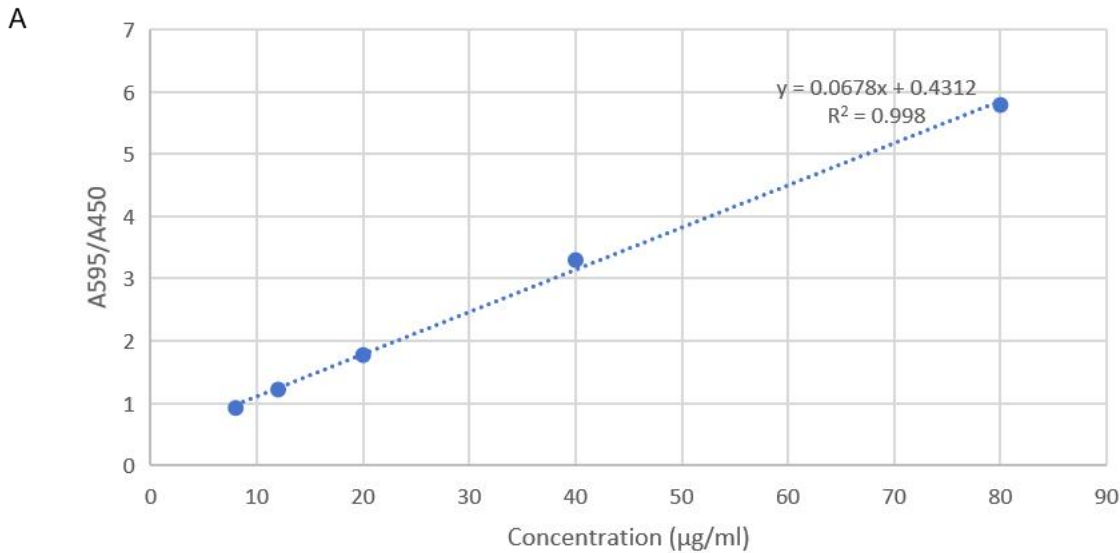
Table 2.6 Information on the antibodies and their dilutions or volumes in different experiments.

Chapter 3 Results

The study aims to map the binding of VEZF1 across the genome of K562 cells, and to explore the interactions with other transcription factors (TFs) and chromatin regulators by CUT&Tag. CUT&Tag requires high activity and specificity of the antibodies, therefore western blotting and immunofluorescence were used to find the most suitable antibodies and their concentration. Considering that all the proteins involved are located in the nucleus, nuclear extracts were prepared for western blotting. To check the specificity of antibodies to VEZF1, nuclear extracts were prepared from both parental K562 cells, and from VEZF1 knockout (VEZF1 KO) K562 cells. The VEZF1 KO cells were created using CRISPR nickase technology, which created small deletions and frame shifts in exon 2, resulting in the loss of full length protein production (Al-Hosni, 2016).

3.1 Preparation and concentration detection of cell nuclear extracts

The concentration of the protein in the nuclear extract was detected by Bradford Assay, a dye-binding assay in which the color of the dye changes in response to different concentrations of protein. The maximum wavelength of absorbance of Coomassie Brilliant Blue G-250 dye shifts from 465 nm to 595 nm when bound to protein. The absorbance can be measured by microplate reader and the protein is quantified with the application of Beer's Law. A series of dilutions of BSA solution were used as calibration. It was found that the A_{595} data points could not be fitted with a straight line, so the A_{595}/A_{450} ratio was used instead. This ratio reduces the interference of EDTA and salt in the nuclear extract and results in a linear calibration plot (Figure 3.1).



B

	Parental	VEZF1 KO
Concentration (mg/ml)	2.08	1.78
Yield (µg)	208	178
Yield per million cells (µg/10 ⁶)	4.03	3.58

Figure 3.1 Determination of the concentration of two nuclear extracts.

Protein concentrations were determined by adding Bradford reagent and reading the absorbance at 595 nm and 450 nm using a microplate reader. A) Calibration curve showing the A_{595}/A_{450} of a BSA dilution series. Data points show the average from three replicates. The best fit straight line was plotted using Microsoft Excel. B) Nuclear extract concentrations from parental K562 cells and VEZF1 knockout K562 cells. For each extract, Bradford assays were carried out with two different nuclear extract dilutions with two replicates of each. The average A_{595}/A_{450} ratio was calculated, and the protein concentration of each nuclear extract was calculated according to the linear equation in the scatter plot.

3.2 Western blotting

3.2.1 Testing the original stock of VEZF1 antibody (AGW-3642-old)

Rabbit anti-VEZF1 polyclonal antibodies were previously raised against a conserved VEZF1 peptide, Ser376-Ala547, that lies in the C-terminal half of the 521 amino acid protein. Serum was collected from two rabbits (3642 and 3645) and IgG was isolated using protein A purification (Dickson et al., 2010). The activity and specificity of the previously prepared VEZF1 antibody stock (AGW-3642-old) was determined by western blotting (Figure 3.2). The nuclear extracts of both parental K562 cells and VEZF1 KO K562 cells were used to locate the bands of VEZF1 and identify any non-specific binding. A commercial SP1

antibody was used as a control, and the presence of the expected 81 kDa band for SP1 confirms the quality of the two nuclear extracts. Several dilutions of the size markers and antibodies were tested to optimise the signal and specificity (Figure 3.2).

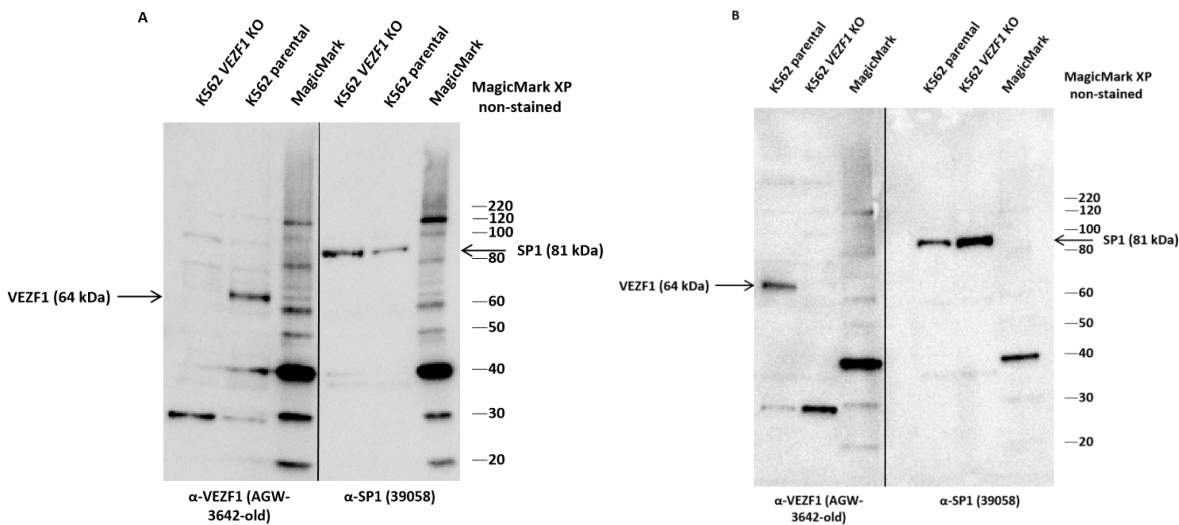


Figure 3.2 Testing the activity and specificity of the stored VEZF1 antibody.

Western blots of nuclear extracts (17.8 μ g) from parental and VEZF1 KO K562 cells were probed with anti-VEZF1 (AGW-3642-old) and anti-SP1. A) The marker was undiluted. Both VEZF1 antibody (AGW-3642-old) and SP1 antibody were diluted 1:5,000. The secondary antibody (anti-rabbit HRP) was diluted 1:2,000. B) The marker was diluted 1:2.5. The VEZF1 antibody was diluted 1:20,000 while the SP1 antibody was diluted 1:10,000. The secondary antibody was diluted 1:10,000.

A 64 kDa band corresponding to the expected size for full length VEZF1 was only found in the nuclear extract from K562 parental cells and not in the VEZF1 KO cells, confirming the identity of this band. The higher concentration of anti-VEZF1 in Figure 3.2A reveals several additional bands. The anti-VEZF1 antibody is a mixture of IgG antibodies rather than a single affinity-purified antibody, which makes it more likely that non-specific bands will be observed. Therefore, a higher dilution factor for the anti-VEZF1 antibody was used in Figure 3.2B, and most of the non-specific bands were eliminated. However, a strong, non-specific band at 28 kDa was still observed in both nuclear extracts probed with anti-VEZF1. The 28 kDa band cannot be a short form of VEZF1, as the frameshift generated by the CRISPR machinery is upstream of the peptide that the antibody was raised against. The results showed that the anti-VEZF1 antibody correctly binds to VEZF1 in the nuclear extracts, but indicates that non-specific binding could be a concern for the CUT&Tag procedure.

3.2.2 Purification and verification of new VEZF1 antibodies

In order to find a prospective VEZF1 antibody with higher activity and specificity, two new batches of anti-VEZF1 were purified from the serum of two different rabbits (3642 and 3645) using protein A-based column chromatography. Bradford assays were used to measure the concentration of the purified VEZF1 antibodies AGW-3642-new and AGW-3645-new (Table 3.1).

	3642 IgG	3645 IgG
Concentration (mg/ml)	2.64	1.01

Table 3.1 Detecting the concentration of new anti-VEZF1 IgG.

Concentration and yield of two types of anti-VEZF1 IgG were determined using Bradford assays as described above and calculated according to the linear equation in the scatter plot.

3.2.3 Verification and comparison of old and newly purified VEZF1 antibodies

To verify and compare the activity and specificity of AGW-3642-old, AGW-3642-new, and AGW-3645-new, western blotting was performed (Figure 3.3).

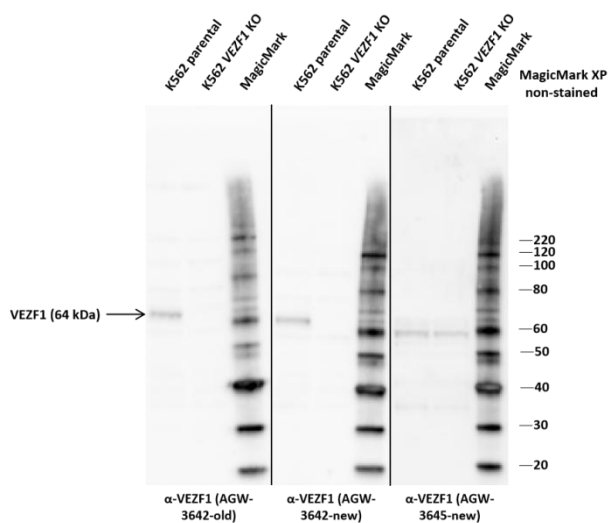


Figure 3.3 Verification and comparison of activity and specificity of three VEZF1 antibodies.

The activity and specificity of old and newly purified VEZF1 antibodies was tested via Western blot method. The arrow points at the bands of VEZF1. The marker was undiluted. All the VEZF1 antibodies were diluted 1:5,000. The secondary antibody was diluted 1:100,000.

It can be seen that the AGW-3645-new preparation did not detect a band of the correct size for VEZF1 (64 kDa), and that there was no difference in intensity of the 60 kDa band between the parental and VEZF1 KO K562 samples. As similar results were obtained in repeated experiments, it was concluded that

AGW-3645-new was not suitable for detecting VEZF1.

Both the old and new preparations of AGW-3642 show specificity for VEZF1, and under these conditions the non-specific band at 28 kDa is not detected. Additional blots with different conditions and new nuclear extract preparations revealed that the AGW-3642-new antibody had fewer non-specific bands than the AGW-3642-old preparations (data not shown). Therefore, AGW-3642-new was chosen to be used in future experiments.

3.2.4 Selection and concentration optimization of commercial antibodies

In addition to the purified AGW-3642 VEZF1 antibody, five commercial antibodies were tested, which were specific to GATA2 (PA1-100, NBP1-82581, ab153820), CTCF (ab188408), and VEZF1 (NBP1-84301-25ul) (Figure 3.4).

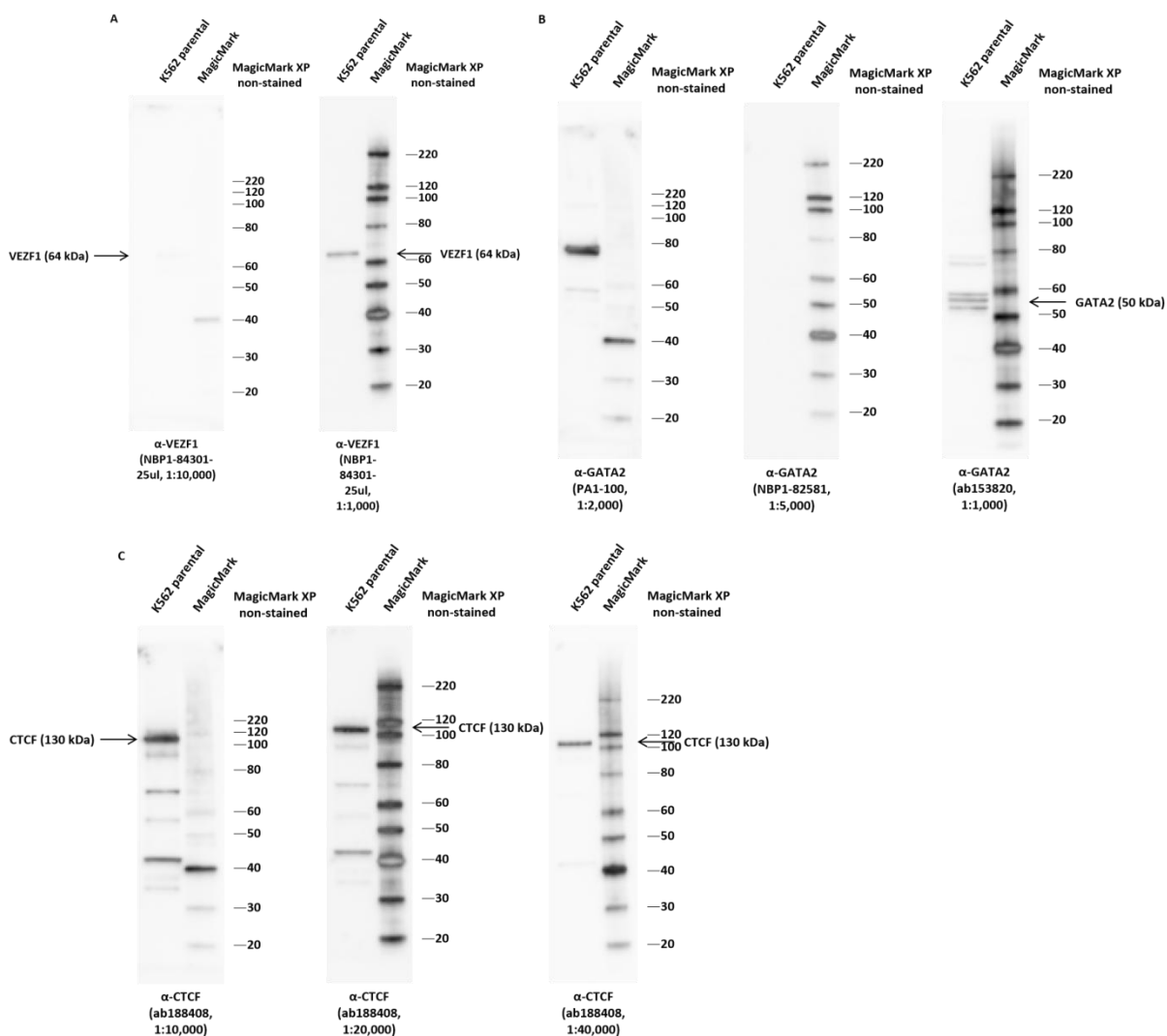


Figure 3.4 Verification and optimization of commercial antibodies.

The activity and specificity of five commercial antibodies was tested via Western blot method. The arrows point at the bands of the target proteins (or where they should present). The marker was undiluted. The secondary antibody was diluted 1:100,000. A)The strips for VEZF antibody (NBP1-84301-25ul). B)The strips for three GATA2 antibodies (PA1-100, NBP1-82581, ab153820). C)The strips for CTCF antibody (ab188408).

PA1-100 was the first GATA2 antibody tested, and it revealed a clear band on the blot when used a a range of antibody dilutions (Figure 3.4 and data not shown). However, the band was around 80 kDa instead of 50 kDa, which is the molecular weight of GATA2 protein (Gordon et al., 1997). Although GATA2 can run between 50 kDa and 60 kDa in different buffer systems, 80 kDa was considered to be too large to be GATA2, indicating that PA1-100 did not bind specifically to GATA2 protein. A second GATA2 antibody (NBP1-82581) was tested. Disappointingly, a trace of the band can only be distinguished using large amounts of antibody and long exposure times. The performance of NBP1-82581 was equally poor in subsequent immunofluorescence tests, as no green fluorescence was observed (data not shown), hence this antibody was abandoned. The last GATA2 antibody tested was ab153820, which gave a triplet of bands of 50-60 kDa on repeated western blots (Figure 3.4B). As this antibody was raised against a large region of GATA2 that is well conserved between other family members, it is possible that one of these bands corresponds to GATA1, which is also expressed in K562 cells and also has a molecular weight of around 50 kDa. Another potential reason for multiple bands could be phosphorylation of GATA2 (Koga et al., 2007), as the phosphorylated protein migrates slightly slower than the non-phosphorylated version. It is also possible that some GATA2 protein was degraded during storage. A lack of time meant that these possibilities could not be investigated further. Although the triple band was not a satisfying result, ab153820 was the only GATA2 antibody that gave bands with the expected molecular weight of 50-60 kDa. Thus, ab153820 was selected for CUT&Tag.

The CTCF antibody showed extremely high activity in western blotting, which resulted in non-specific bands when used at dilutions of 1/10,000 and 1/20,000. A range of antibody dilutions were tested in order to find the most suitable dilution factor with fewest non-specific bands. A dilution factor of 1 in 40,000 was found to eliminate the non-specific bands, leaving only a band at the expected size for CTCF of 130 kDa (Figure 3.4C).

A commercial antibody against VEZF1 (NBP1-84301) was tested as a comparison with the AGW-3642-new antibody that was presented in Figure 3.4A. When used at a dilution of 1/1000 it gave a single band of the expected size of 64 kDa.

3.3 Immunofluorescence

It is recommended that the ideal antibody concentration for CUT&Tag is that same as that used for immunofluorescence (Meers et al., 2019). Therefore, immunofluorescence was performed to optimise antibody concentration and to check that the signal for each protein was specifically localised to the nucleus. Two dyes were used as controls to locate the cells in the images. Hoechst 33342 (2'-[4-ethoxyphenyl]-5-[4-methyl-1-piperazinyl]-2,5'-bi-1H-benzimidazole trihydrochloride trihydrate) is a cell-permeable DNA stain that binds preferentially to adenine-thymine (A-T) regions of DNA and can be detected with a fluorescent microscope at $E_m = 460-490$ nm (blue). Phalloidin-iFluor 555 Reagent binds to actin filaments, also known as F-actin. Phalloidin-iFluor 555 can be detected with a fluorescent microscope at $E_m = 574$ nm (red).

In the early stage, multiple concentrations of antibodies and dyes were tested, but bright fluorescence was not observed even the concentration have exceeded the maximum of the recommended range from the manufacturer. The issue was solved by using a fluorescence EVOS microscope which is designed for well plates. Finally the clear images of fluorescence were acquired (Figure 3.5).

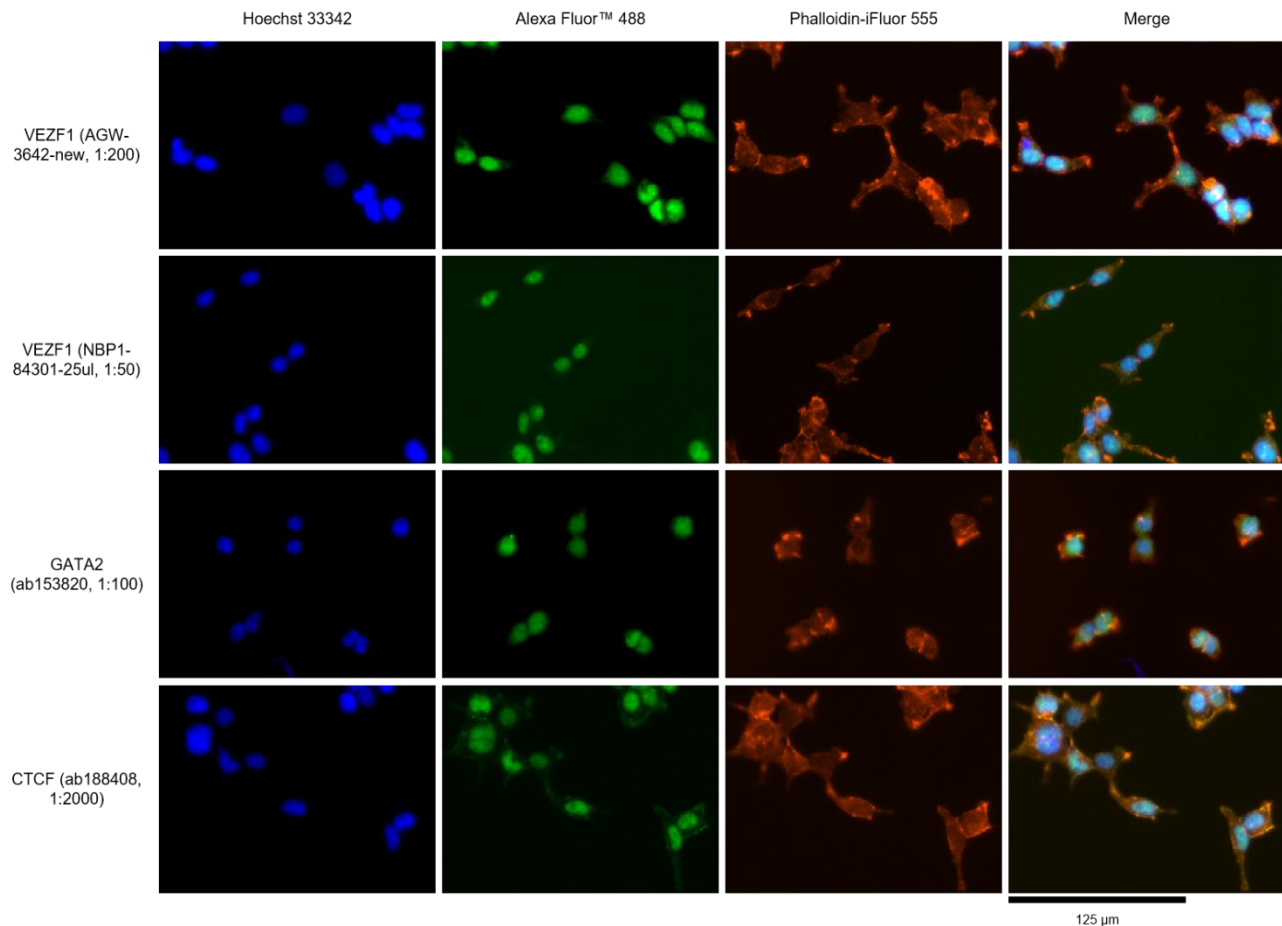


Figure 3.5 Test of the activity and specificity of antibodies by immunofluorescence.

HEK cells were separately incubated with 4 primary antibodies and immunostained in 3 colours: staining for Hoechst 33342 (blue); Phalloidin-iFluor 555 (red); and target proteins (green), was performed. The secondary antibody (A-11008) was diluted 1:500.

Cell bodies are indicated by the red actin staining, and nuclei are indicated by blue Hoechst staining of DNA. All the four primary antibodies against VEZF1, GATA2 and CTCF show strong nuclear staining at the concentrations used. However, the general green tinge to the whole field of view was higher with the commercial VEZF1 antibody (NBP1-8430) than with the lab-purified VEZF1 antibody (AGW-3642-new). Although NBP1-84301-25ul showed better results than AGW-3642-new in western blotting, the latter was applied subsequently in CUT&Tag.

3.4 Use of CUT&Tag methodology to map the distribution of VEZF1 and other factors in K562 cells

3.4.1 Preparation of CUT&Tag samples for next generation sequencing

To map the distribution of VEZF1 in K562 cells, and investigate the relationship between VEZF1, GATA2 and CTCF, CUT&Tag was performed using the antibodies that were optimised in the previous section. CUT&Tag was also performed with well validated commercial antibodies to the histone modifications H3K4me1, H3K4me3 and H3K27ac. The CUT&Tag procedure first involves immobilising living cells on magnetic beads coated with Concanavalin A, which binds to glycoproteins on the cell membrane (Meers et al., 2019). Cells are permeabilised with digitonin, which allows the primary antibodies to enter the cell nucleus and bind to their targets. Incubation with an appropriate secondary antibody increases the number of protein A binding sites at each target. The cells are then incubated with a protein A-tagged Tn5 transposome that has been pre-loaded with DNA adapters. The transposome is recruited to the target sites, where it integrates the adapter sequences into the DNA, a process that has been called tagmentation. This generates DNA fragments tagged with adapter sequences that can be amplified by PCR and analysed by next generation sequencing.

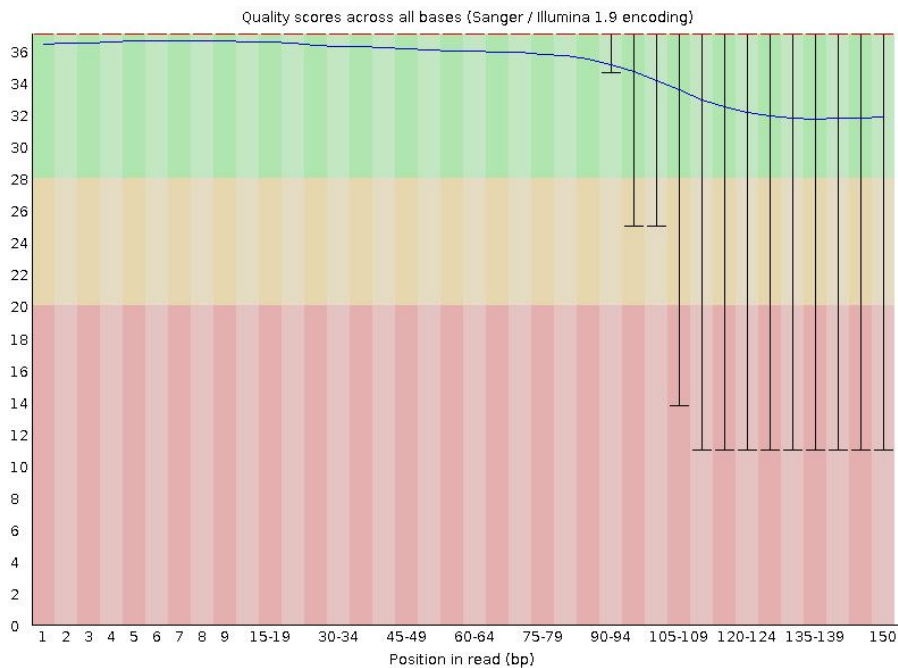
In order to improve the reliability of the data, several CUT&Tag replicates were performed for each antibody. In addition, three different dilutions of the CTCF antibody were used. This is because the western blotting revealed that the CTCF antibody was very active and showed non specific binding unless a very high dilution of 1 in 40,000 was used. We were concerned that using too much CTCF antibody for CUT&Tag would result in higher background or more false positive peaks.

3.4.2 Quality control of NGS data

Samples were subject to 150 bp paired-end sequencing (Novagen). The number of reads for most samples was between 15.9-38 million, although one sample only had 5.9 million reads. The quality of the raw data was checked and controlled before further bioinformatic analysis was carried out. Sample contamination and instrument errors can result in low quality data, which would reduce the credibility of the results of analysis. Quality control ensures that analysis are performed on high quality data, thus

enabling robust conclusions to be drawn.

A



B

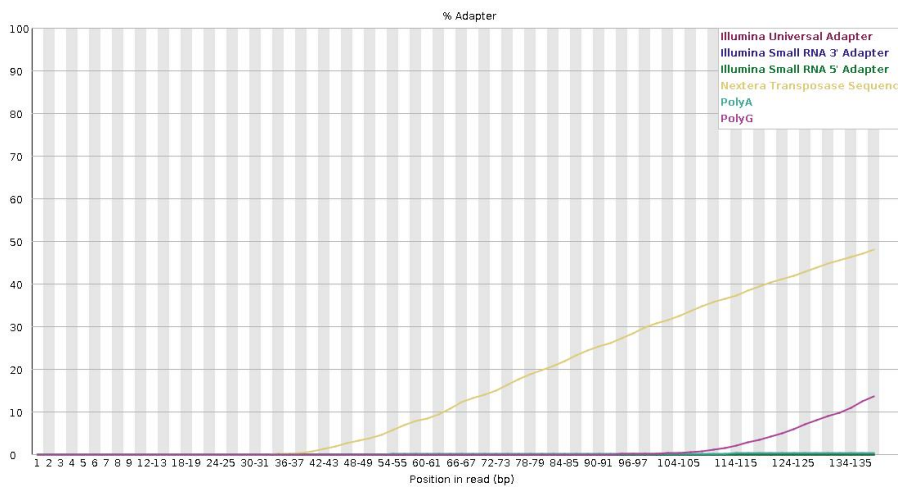


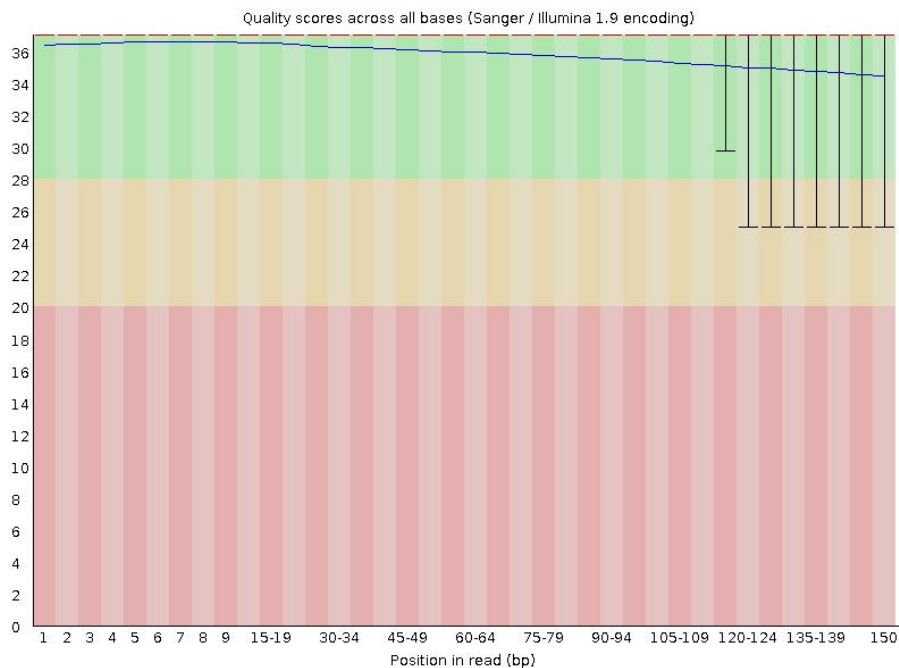
Figure 3.6 FastQC on raw data.

A) Per base sequence quality of the raw data. The blue line represents the mean quality score of each base, and the black whiskers represent the 10% and 90% points. The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red) (Andrews, 2010). B) Adapter content of the raw data. Lines with different colours represents the proportion of different adapter sequence.

The Phred score of the raw data decreased substantially for bases more than 100 bp into the read (Figure 3.6A), which was possibly due to the degradation of fluorophores and the existence of strands in the cluster not being elongated (Andrews, 2010; hbctraining.github.io). In addition, adapter sequences

were detected towards the 3' ends (Figure 3.6B). To improve the quality of the data, the Cutadapt tool was applied to trim off the adapter sequences. After trimming, FastQC was run again and showed that the quality of the data had been improved (Figure 3.7).

A



B

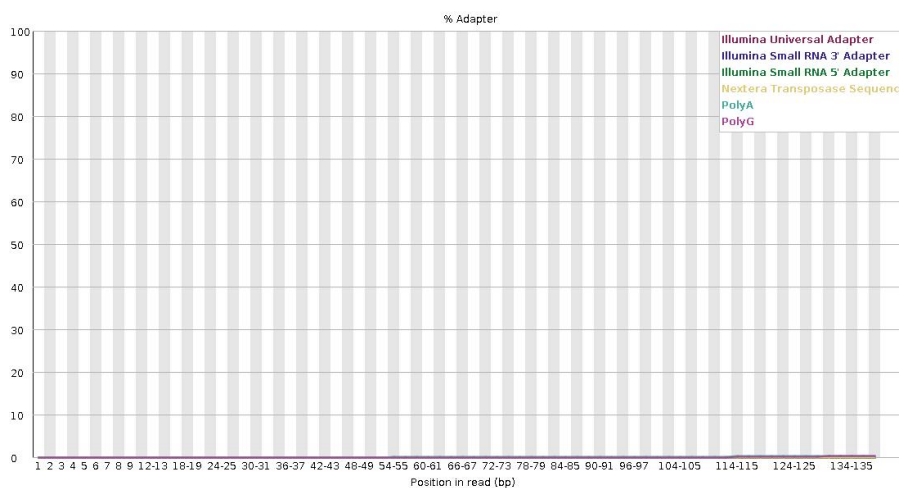


Figure 3.7 FastQC after trimming.

A) Per base sequence quality of the data after trimming. The mean quality score of each base was increased, especially for bases more than 100 bp, and the black whiskers were shortened. B) Adapter content was not detected after trimming.

3.4.3 Alignment to the genome and visualisation of CUT&Tag data

After the quality of data was guaranteed, Bowtie2 was applied to align the reads to the Hg19 genome,

generating a bam file. Bowtie2 is a widely used tool in bioinformatics that aligns sequencing reads to large genomes with high effectiveness and efficiency. It takes account of the Phred score of each base during alignment, and ensures that the read-mate pairs are aligned concordantly wherever possible. 87% of reads were aligned concordantly to the Hg19 genome.

To visualise the CUT&Tag data, bam files were converted into bigwig files for display on the UCSC genome browser. bigwig tracks for each replicate were examined visually to check for high enrichment and low background, before datasets were taken forward for subsequent analysis. Figure 3.9 and Figure 3.10 show the individual replicates for the VEZF1 samples. One of the datasets failed the visual checking step: H3K27ac_2. This sample had low enrichment at expected peak sites, and high background signal between peaks, so it was eliminated from subsequent analysis (data not shown).

Comparison of the bigwig tracks for CUT&Tag performed with different dilutions of CTCF antibody did not reveal any obvious differences. The locations and relative heights of the peaks, and the level of background noise did not change whether 0.1 µg, 0.5 µg or 1 µg of CTCF antibody were used (data not shown). Therefore, data from all the CTCF samples was carried forward into the subsequent analysis.

3.4.4 Investigation into duplicate reads

The quality control analyses performed by FASTQC indicated a higher than normal level of duplicate reads (data not shown). The sequencing of duplicate reads during next generation sequencing can be an indicator of biased PCR amplification during the library preparation stage, and could mean that the output data is not representative of the initial sample. High levels of duplicate reads can be higher than normal when the initial sample size is very small, if too many PCR cycles are used to amplify the library, and/or when a small sample is sequenced to a very large depth. However, Cut&Tag datasets often contain higher than normal levels of duplication due to the lower complexity of the initial sample as the background signal is extremely low, and this is not typically considered to be a problem during data analysis (Ye Zheng et al., 2020).

To quantify the level of sequence duplication in the samples prepared in this project, the “Mark Duplicates” from the Picard tool set was used to interrogate the bam alignment files (Broad Institute of MIT and Harvard, 2014). Table 3.2 shows that the fraction of each library that consists of duplicates is

greater than 20% for all except three samples. It is notable that the three samples with lower rates of duplication, H3K4me3_1, H3K4me1_2 and H3K4me1_3, were generated from histone modifications rather than transcription factors, so are likely to have more complex initial libraries. They were also sequenced to a slightly lower level, with libraries of 9 million reads or less. For CUT&Tag, typically 5 million paired ends are used, whereas our samples range from 2.9 to 18.8 million reads.

Data set	READ PAIRS EXAMINED	PERCENT DUPLICATION	ESTIMATED LIBRARY SIZE
VEZF1_AGW_1	17630316	0.891337	1917429
VEZF1_AGW_2	15496703	0.730311	4306475
VEZF1_3_1	16688144	0.883227	1951453
VEZF1_3_2	18233947	0.906663	1703671
GATA2_1	11883082	0.838596	1920326
GATA2_2	13489176	0.831807	2276117
GATA2_3	12429027	0.868765	1631994
CTCF_2	17024388	0.691235	5519135
CTCF_01_1	12221356	0.750607	3115482
CTCF_01_2	11811291	0.860389	1651441
CTCF_05_1	14426330	0.659854	5264481
CTCF_05_2	13097464	0.698736	4134322
H3K4me3_1	2929768	0.117286	12077189
H3K4me3_2	9131616	0.211017	19223254
H3K4me1_1	18884340	0.284396	27401444
H3K4me1_2	7900502	0.148257	25299686
H3K4me1_3	8528238	0.152032	26512525
H3K27ac_1	17452611	0.551876	9255612
H3K27ac_2	15735309	0.638769	6195333
H3K27ac_3_1	15267700	0.57043	7638386
H3K27ac_3_2	13447721	0.576989	6564902

Table 3.2 Duplicate reads in the CUT&Tag data sets.

Each replicate was analysed using the Mark Duplicates tool from the Picard set of tools. The table shows the total number of aligned read pairs in each bam file, the fraction of reads that are presents as duplicates, and the estimated library size after removal of duplicates.

3.4.5 Peak finding

The purpose of peak finding is to find transcription factor binding and histone modification sites. Model-based Analysis of ChIP-Seq (MACS) is a popular tool used in exploring genome-wide protein-DNA interactions. Here the callpeak function of MACS2 was utilised to find peaks from the alignment results. Removal of duplicates prior to generating the bigwig display files did not alter the location or height of the vast majority of the peaks (Figure 3.8). Some smaller peaks were not called after duplicate removal,

however. For example, the peak at around chr16: 232,000 in Figure 3.8 is not present in track “vezf1_3_1 markduplicates”

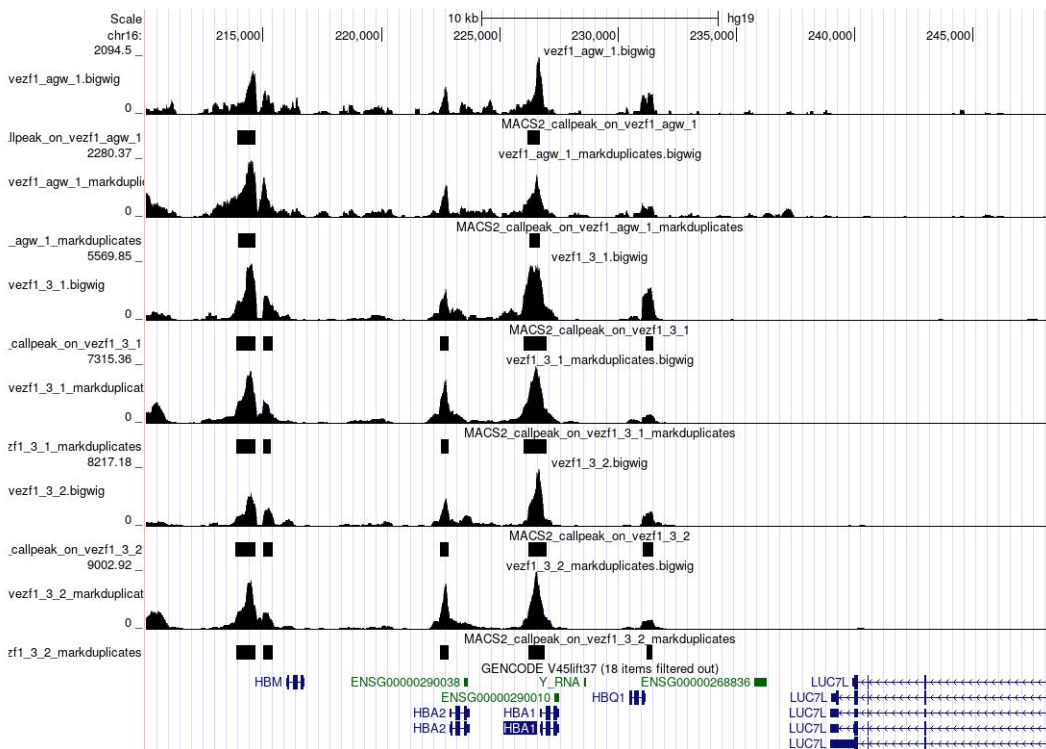


Figure 3.8 VEZF1 bigwig tracks (duplicates removed) with MACS2 at *HBA1* locus.

Two bigwig tracks for each of three VEZF1 replicates (agw_1, 3_1 and 3_2) are shown at the *hemoglobin subunit alpha 1 (HBA1)* locus in K562 cells. The upper track for each replicate includes all the duplicate reads, and the lower track (markduplicates) has the duplicates removed. Peaks called by MACS2 are shown as blocks beneath each track.

At first, MACS2 callpeak was carried out on each replicate separately (duplicate reads included), and then bedtools Intersect intervals was used in order to identify peaks that were present in all replicates for a particular antibody. This approach is very robust, as it eliminates peaks that are not found in all replicates. However, it runs the risk of removing smaller peaks that represent true binding to weaker sites, and which may be important for our subsequent analysis. Therefore, samtools merge was applied to merge the bam alignment files from replicate samples, and MACS2 callpeak was performed on the merged bam file. This means that the peaks were identified from the total data set for each antibody, resulting in maximum sensitivity.

Figure 3.9 illustrates this process for the VEZF1 replicates. The bigwig track and MACS2 peaks are shown for each of the four VEZF1 replicates, and for the merged bam file at the *haemoglobin subunit alpha 1*

(*HBA1*) locus. The central VEZF1 peak at *HBA1* is identified in all four replicates and in the merged bam file. However, the peaks at *HBA2* and *HBQ1* were not called in replicate VEZF1_agw_1, presumably because their enrichment compared to the surrounding background signal was not strong enough. In the merged bam file, the three peaks at *HBA2*, *HBA1* and *HBQ1* are clearly enriched above the background signal, and all three peaks were called by the MACS2 algorithm.

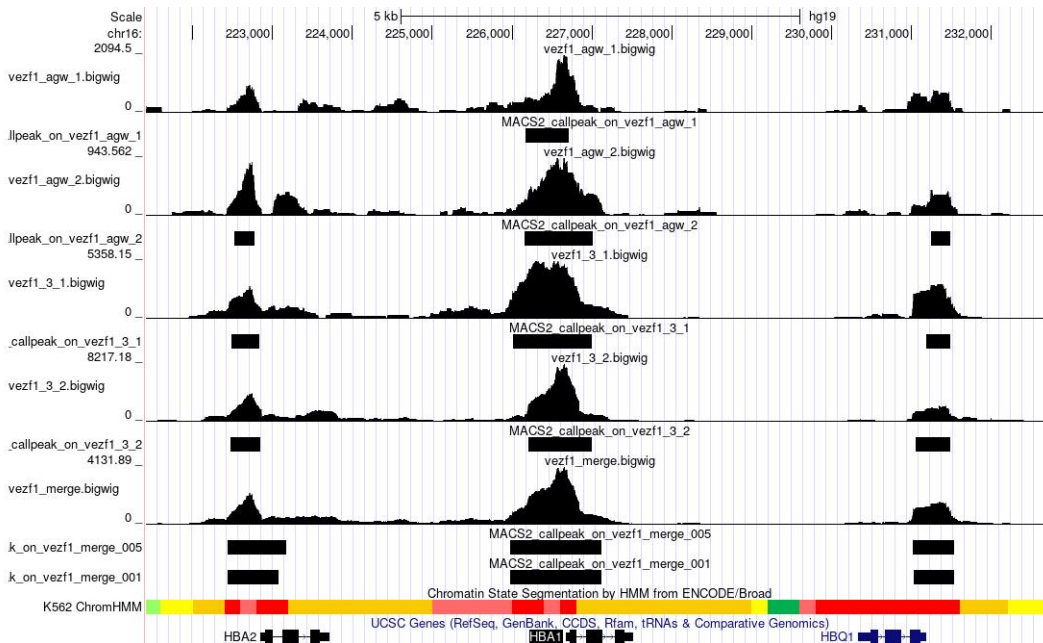


Figure 3.9 VEZF1 bigwig tracks (merged) with MACS2 at *HBA1* locus.

Individual and merged VEZF1 bigwig tracks with enriched peaks identified by MACS2 at the *HBA1* locus in K562 cells. From top to bottom, four tracks for individual VEZF1 replicates (agw_1, agw_2, 3_1 and 3_2) and their peaks called by MACS2, the track for the merged VEZF1 dataset and its peak calling using a minimum FDR cutoff of 0.05 or 0.01 by MACS2 are shown.

The increased sensitivity of peak calling when using the merged bam files could also result in identification of false positive peaks, however. False positives are known to be an issue when using MACS2 to call peaks from CUT&tag data, as the low background can mean that a random accumulation of reads is called as a peak by the algorithm. This is illustrated in figure 3.10, which shows the VEZF1 replicates and merged bam tracks in the region of the *TAL1* locus. MACS2 has called a false positive peak downstream of the *FOXD1* gene, which does not correspond to any visible enrichment in the bigwig tracks (Meers et al., 2019). In order to reduce the number of false positive peaks in the background, a more stringent MACS2 strategy was used, which is to adjust the minimum false discovery rate (FDR, q-value) cutoff for peak detection from 0.05 to 0.01. In this way the false positive peaks were minimised, and the credibility of the data was improved (Figure 3.10).

The disadvantage of this merge strategy is that combining unequal numbers of reads from each replicate could lead to bias in the final file. To improve the analysis pipeline in the future, equal numbers of reads should be merged from each replicate after duplicate reads have been removed.

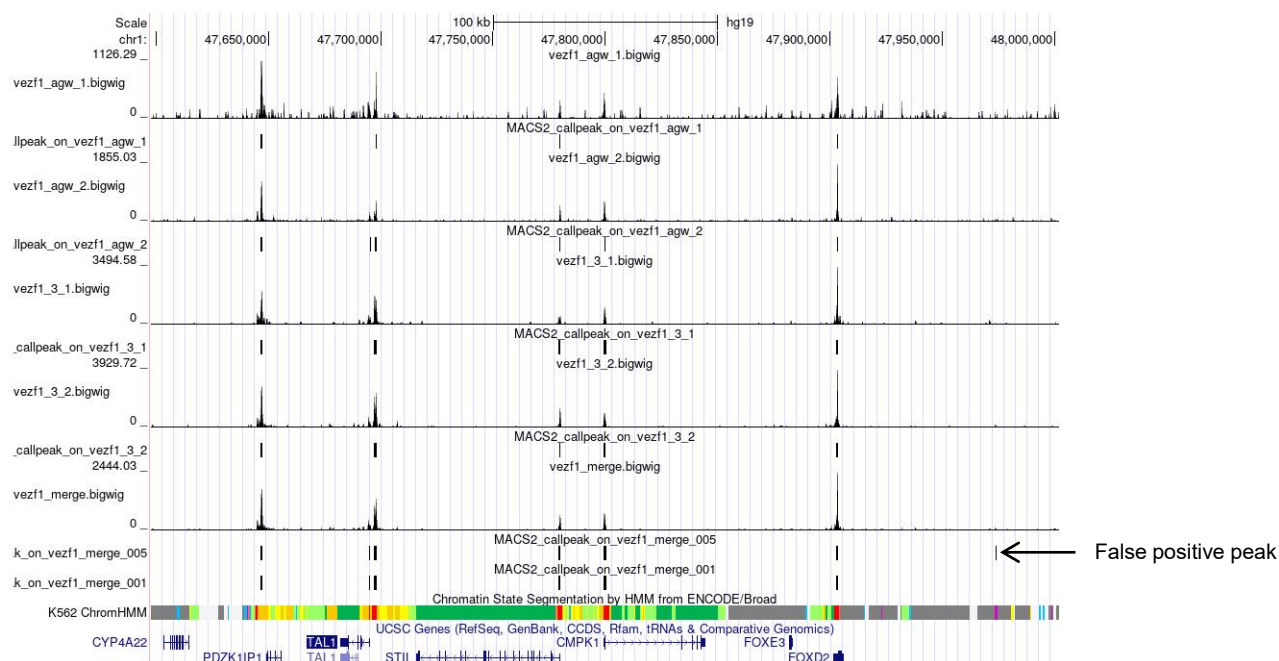


Figure 3.10 VEZF1 bigwig tracks with MACS2 at *TAL1* locus.

Individual and merged VEZF1 bigwig tracks with enriched peaks identified by MACS2 at *TAL1* locus. From top to bottom, four tracks for individual VEZF1 replicates (agw_1, agw_2, 3_1 and 3_2) and their peaks called by MACS2, the track for the merged VEZF1 dataset and its peak calling using a minimum FDR cutoff of 0.05 or 0.01 by MACS2 are shown. The arrow points out the false positive peak that occurs at a minimum FDR of 0.05.

From a genome-wide perspective, the total number of peaks called for the individual VEZF1 replicates were 11,832, 9001, 9667 and 6944. The number of peaks that were common to all four replicates was 4235, indicating that many peaks were lost by using this robust filtering approach. In contrast, 17,605 peaks were identified from the merged bam file when using FDR of 0.05, and the more stringent setting of FDR 0.01 resulted in 14,320 peaks.

It should be mentioned that the Henikoff lab have produced an alternative peak calling algorithm for CUT&Tag data, SEACR, which should reduce false positive peaks (Meers et al., 2019). However, it requires an IgG control track for optimum results, which we did not generate. It is possible to run the software using settings that do not require a control sample. However, even the most stringent settings resulted in over 100,000 peaks, most of which were clearly false positives when compared to the bigwig tracks (data not shown).

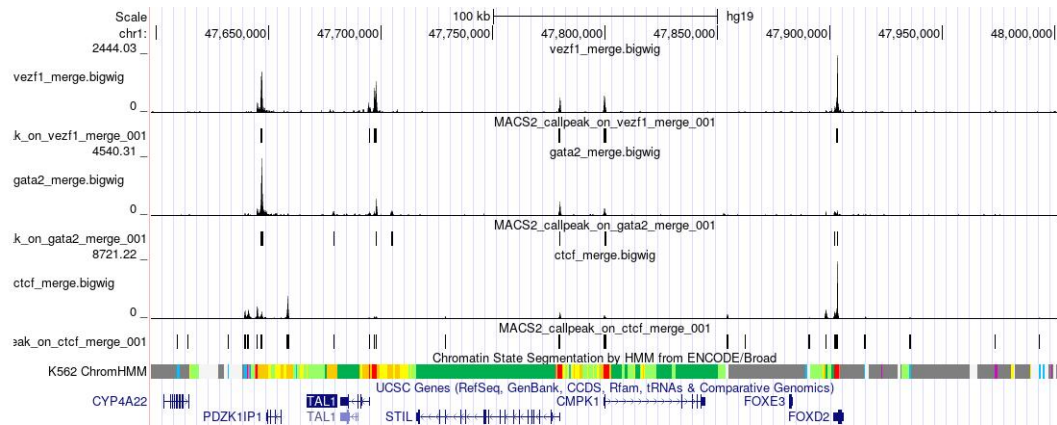
Based on these comparisons, it was concluded that the following approach would be used for all datasets: 1) visual checking of the bigwig file for each replicate 2) merging the bam files from all high quality replicates, and 3) peak calling using MACS2 on the merged bam file with an FDR of 0.01, as this strategy results in high sensitivity for identifying peaks while reducing the number of false positive peaks.

3.4.6 Enrichment of VEZF1, GATA2, CTCF and histone modifications at the *TAL1* locus

Figure 3.11 shows the tracks for GATA2, CTCF and histone modifications in the vicinity of the *TAL1* locus. Peaks for VEZF1, GATA2, H3K4me1, H3K4me3 and H3K27ac can be observed at the transcription start sites for *TAL1*, *STIL1*, *CMPK1* and *FOXD2*. No enrichment is observed at *CYP4A22*, *PDZK1IP1* and *FOXE3*. RNAseq data from the ENCODE project, available on the UCSC genome browser, reveals that *TAL1*, *STIL1*, *CMPK1* and *FOXD2* are all expressed in K562 cells, whereas *CYP4A22* and *FOXE3* are not (data not shown). *PDZK1IP1* is expressed at low levels. The highest GATA2 peak is located at 47,647,000, and corresponds to the +51 kb *TAL1* enhancer (Zhou et al., 2013). VEZF1 and the three histone modifications also have peaks there. This data confirms that VEZF1, GATA2 and the three histone modifications are enriched at active promoters and enhancers in K562 cells.

The CTCF peaks do not follow the same pattern as the other factors, with only two major peaks visible in the view shown in Figure 3.11. Notably, the major CTCF peak located at 47,658,600 is a known CTCF binding site that regulates looping interactions at the *TAL1* locus. This is consistent with the role of CTCF in regulating genome architecture via chromatin loop formation, rather than participating in the assembly of protein complexes at gene promoters and enhancers (Zhou et al., 2013).

A



B

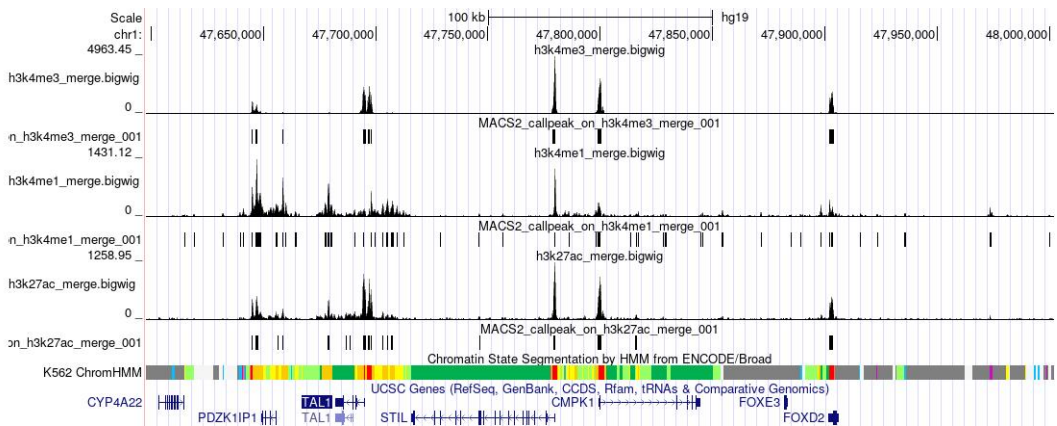


Figure 3.11 Bigwig tracks with MACS2 at *TAL1* locus.

A) Individual and merged bigwig tracks of VEZF1, GATA2, and CTCF with peaks identified by MACS2 at the *TAL1* locus in K562 cells. From top to bottom, CUT&Tag tracks for merged VEZF1, merged GATA2, merged CTCF and their peak calling by MACS2 are shown. B) Individual and merged bigwig tracks of H3K4me3, H3K4me1, and H3K27ac with enriched peaks identified by MACS2 at *TAL1* locus. From top to bottom, our tracks for merged H3K4me3, merged H3K4me1, merged H3K27ac and their peak calling by MACS2 are shown.

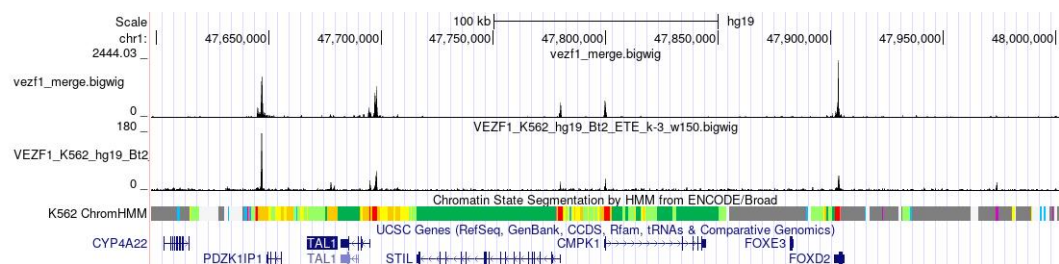
3.4.7 Comparison of CUT&Tag data with ChIP-seq data

One of the objectives of this project was to compare CUT&Tag data with ChIP-Seq data previously generated by the AGW lab. For VEZF1, the peaks show a very similar profile for the ChIP-seq data compared to the CUT&Tag data (Figure 3.12A). The VEZF1 CUT&Tag track had much lower background than the ChIP-Seq track, however, which could result in fewer false positive peaks. The widths of the peaks were generally similar, but tended to be a bit wider for the CUT&Tag data than for the ChIP-seq

data.

The histone modification tracks generated by CUT&Tag showed very similar profiles to previous West lab data generated by both cross-linking ChIP (X-CHIP) and native ChIP (N-CHIP) (Figure 3.12B). The y axis represents the normalised reads per kilobase per million reads in the whole sample (RPKM). The normalised RPKM peaks generated by CUT&Tag are higher than those from ChIP-seq because the background signal across the genome is much lower. Similar peak profiles were also observed for data generated as part of the ENCODE project.

A



B

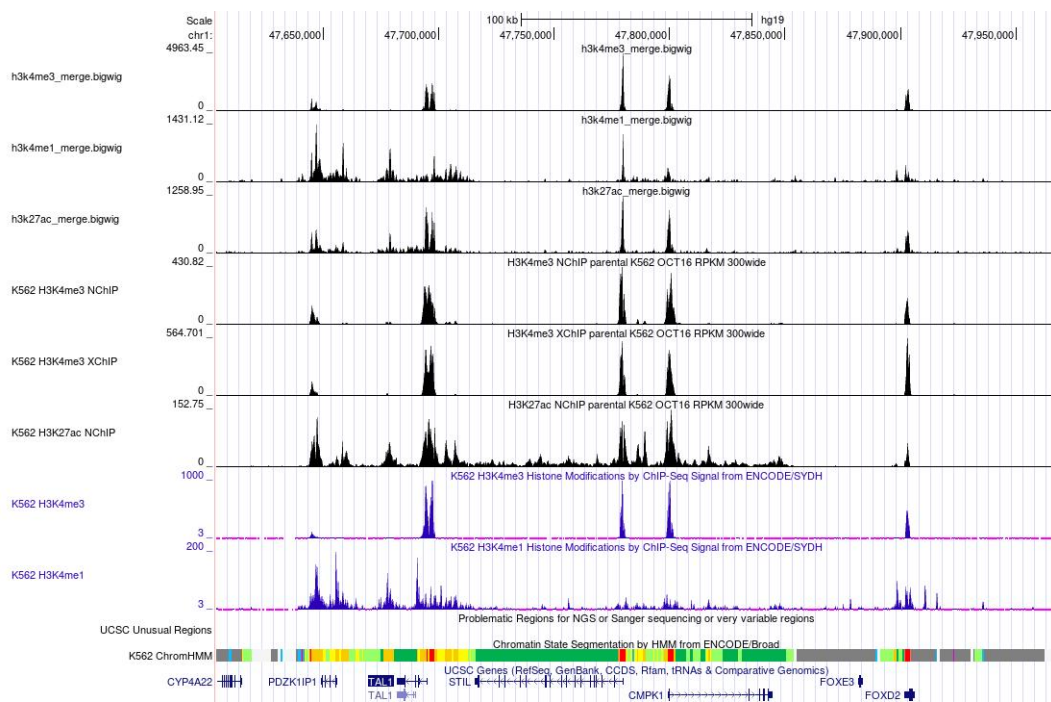


Figure 3.12 Comparison between CUT&Tag data and ChIP-Seq data.

A) Comparison between VEZF1 tracks from our CUT&Tag data and unpublished data from Adam West lab at *TAL1* locus. From top to bottom, our track for merged VEZF1 bigwig by CUT&Tag (K562), track from Adam West lab for VEZF1 (K562) by ChIP-seq are shown. B) Comparison between histone modification tracks from our CUT&Tag data and unpublished data from the Adam West lab at *TAL1* locus. NChIP represents native ChIP-seq, where native chromatin that was treated with micrococcal nuclease to

generate a mix of mono-, di-, and tri- nucleosomes. XChIP represents cross-linking ChIP-seq, where cells were cross-linked with formaldehyde and chromatin was sonicated to 100-600 bp fragments. From top to bottom, our tracks for merged H3K4me3 bigwig by CUT&Tag (K562), merged H3K4me1 bigwig by CUT&Tag (K562), merged H3K27ac bigwig by CUT&Tag (K562), tracks from Adam West lab for H3K4me3 by NChIP (K562), H3K4me3 by XChIP (K562), H3K27ac by NChIP (K562), ENCODE/SYDH histone tracks for H3K4me3 by ChIP-Seq (K562), ENCODE/SYDH H3K4me1 by ChIP-Seq (K562) are shown.

3.4.8 Analysis of peaks

Active promoters and enhancers were defined using chromatin state signatures for K562 cells generated by ChromHMM (Ernst and Kellis, 2010; Ernst et al., 2011). ChromHMM is an open source software package that generates a genome-wide annotation for each cell type by using a multivariate hidden Markov model (HMM) to analyze the marks of chromatin-state signatures (Ernst and Kellis, 2010; Ernst et al., 2011).

The number of VEZF1 peaks overlapping with active promoters and enhancers was quantified, and this data revealed that 70% of VEZF1 peaks are located at promoters, and 14% are located at enhancers (Figure 3.13). There was strong co-localisation of VEZF1 and GATA2 at promoters, with 78% of promoter-associated VEZF1 peaks co-localising with GATA2 peaks, and 80% of promoter-associated GATA2 peaks coinciding with VEZF1 peaks. At enhancers, most (64%) VEZF1 peaks co-localised with GATA2, but only 23% of GATA2-associated enhancers also had VEZF1 peaks (Figure 3.13).

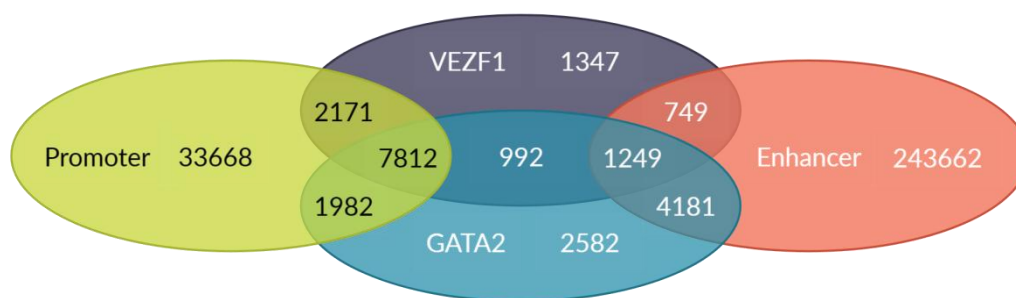


Figure 3.13 Association between VEZF1 peaks, GATA2 peaks, promoters, and enhancers.

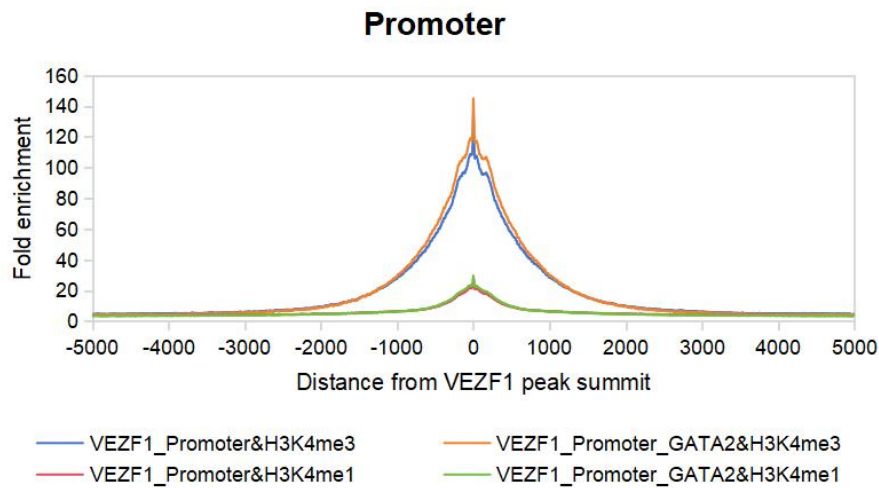
Venn diagram illustrating the number of VEZF1 and GATA2 peaks that overlap with active promoters or enhancers as defined by ChromHMM in K562 cells.

The chromatin features of VEZF1 peaks associated with promoter or enhancer regions were investigated by comparing the enrichment of different histone modifications using the ChIP-Cor tool (Figure 12).

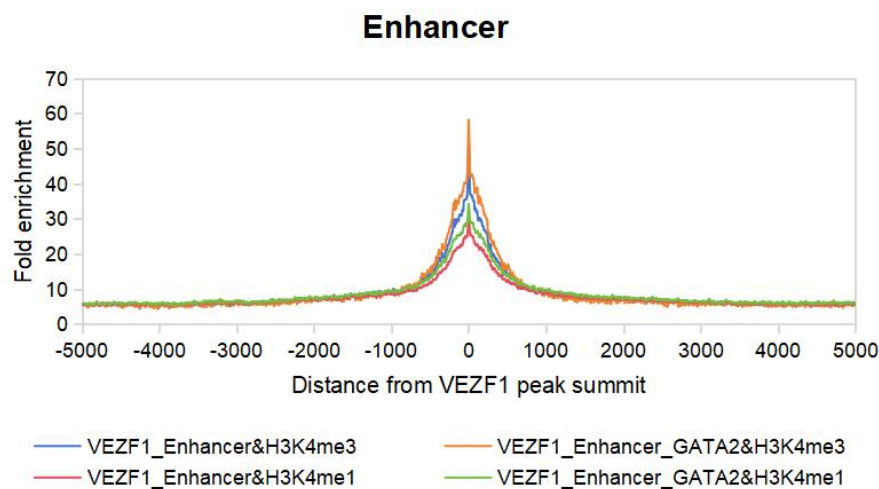
ChIP-Cor generates a positional correlation graph that shows average histone modification enrichment around the summits of a set of peaks or other genomic features.

Analysis of the VEZF1-associated promoters showed that H3K4me3, a key mark of active promoters, was much more highly enriched than H3K4me1 (Figure 3.14A). The level of H3K4me3 was substantially reduced at VEZF1-associated enhancers, although still higher than H3K4me1 (Figure 3.14B). H3K27ac was also more highly enriched at VEZF1-associated promoters than enhancers (Figure 3.14C). These profiles did not change when peaks were filtered to only include those where both VEZF1 and GATA2 were co-localised. This data shows that filtering for GATA2 does not change the overall characteristics of the VEZF1-associated promoters and enhancers.

A



B



C

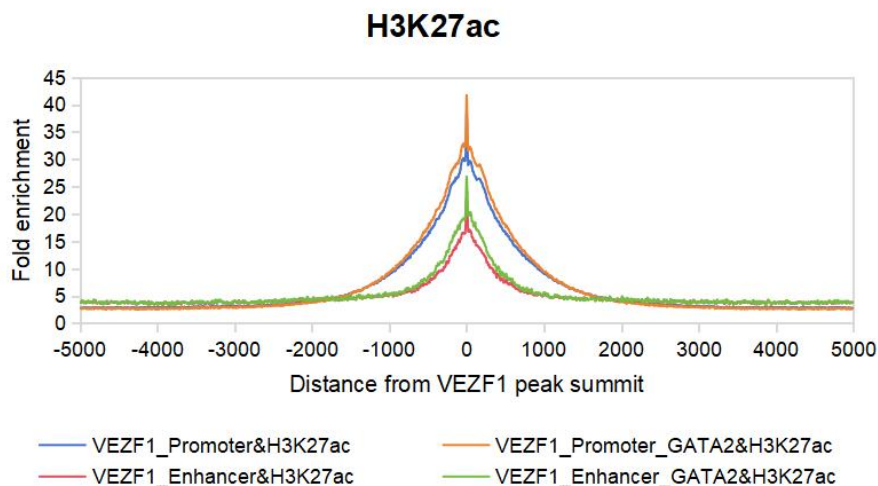


Figure 3.14 Positional correlation graphs of histone modification enrichment at VEZF1 peaks.

A) Mean read density profiles of H3K4me1 and H3K4me3 within ± 5 kb of promoter-associated VEZF1 peak summits. B) Mean read density profiles of H3K4me1 and H3K4me3 within ± 5 kb of enhancer-associated VEZF1 peak summits. C) Mean read density profiles of H3K27ac within ± 5 kb of promoter- or enhancer-associated VEZF1 peak summits. In each graph, profiles either represent data from all VEZF1 peaks at promoters/enhancers, or only data from VEZF1 peaks that also overlap with

GATA2 peaks.

3.4.9 Analysis of DNA sequence motifs

Motif discovery tools such MEME-ChIP can be used to analyse DNA sequences in the vicinity of transcription factor binding sites in order to reveal statistically enriched sequence motifs (Ma et al., 2014). These motifs are likely to represent the consensus DNA binding sequence(s) of the transcription factor or its associated binding partners, and can be used subsequently to analyse DNA binding in more depth. Previous work has shown that VEZF1 preferentially binds to homopolymeric strings of dG:dC *in vitro*, and that this sequence motif is enriched at strong promoter-associated VEZF1 binding sites *in vivo* (Low, 2013). Motif analysis also indicated that VEZF1 binds to weaker GGGGNGGGG motifs *in vivo*, and it was hypothesized that other factors such as GATA1 enable VEZF1 to bind these weaker sites. To investigate whether our CUT&Tag data supports this hypothesis, motif discovery was carried out on the different categories of VEZF1 peaks (Table 3.3).














Category	Rank	Motif	Logo	E-value
All VEZF1	1st	CYCCDCCC		2.2e-296
	3rd	CCDCCKCC		4.3e-070
	6th	CCCGCCYY		1.3e-038
VEZF1&Promoter	1st	CCCGCCYY		7.1e-265
	3rd	CYCCDCCC		6.7e-110
VEZF1&Enhancer	1st	CCMCRCCC		6.8e-027
VEZF1&GATA2	1st	RGGCGGGR		4.3e-229
	3rd	GGMGGAR		3.5e-071
	7th	GGGWGGGR		1.8e-027
VEZF1&Promoter &GATA2	1st	RGGCGGGR		3.6e-212
	3rd	CTCCDCCY		4.9e-066
	8th	CMCRCCCC		1.1e-024
VEZF1&Enhancer& GATA2	2nd	CCMCRCCC		1.5e-014

Table 3.3 VEZF1-like motifs enriched at CUT&Tag peaks.

VEZF1 and GATA2 peaks were identified from aligned reads using MACS2. VEZF1 peaks were then categorized as to whether they overlap with GATA2 peaks and/or promoters or enhancers. For each category of VEZF1 peak, MEME-ChIP was used to identify enriched sequence motifs in the 200 bp immediately surrounding the peak summits. The resulting motifs were ranked according to E-value, which is the motif P value times the number of candidate motifs tested. The table shows VEZF1-like consensus motifs that were enriched in each category, their rank, logo and E value.

Motifs with similarity to the VEZF1 consensus of 9(dG.dC) or a degenerate consensus of GGGNGGGG are shown in Table 3.3. It can be seen that a homopolymeric dG.dC consensus was not identified in any of the sets of VEZF1 peaks. Motifs similar to the GGGCGGGG (or its complement CCCGCCC) were the most highly enriched motif in each VEZF1 category except one. The exception is at enhancers where VEZF1 co-localises with GATA2, where the GATA motif was most highly enriched and the CCCGCCC motif was ranked 2nd. Lower ranking motifs that do not fit the GGGNGGGG pattern, such as CCDCCKCC, GGMGGAR and CTCCDCCY are unlikely to be VEZF1 binding sites, and most likely represent binding sites for other transcription factors (Low, 2013).

Although there are subtle differences in the most highly enriched VEZF1-like motifs in the different categories, there is no consistent evidence to suggest that the consensus binding motif for VEZF1 is different at promoters versus enhancers, or that the VEZF1 motif is altered at sites that co-localise with GATA2. Therefore, this data supports the hypothesis that VEZF1 can bind to weak GGGCGGGG sites *in vivo*. However, it does not provide evidence for a specific role for GATA2 in promoting VEZF1 binding to these sites, either at promoters or enhancers, but given that most VEZF1 sites co-localise with GATA2 in these cells, a role for GATA2 cannot be ruled out.

Analysis of other motifs that are enriched in the different peak subsets revealed the sequence motif for the GATA family: TTATCW (the complement of which is WGATAA) (Table 3.4). This GATA motif was identified in the all-VEZF1 and all-GATA2 peak categories, and at VEZF1 and GATA2 peaks at enhancers. This is consistent with the previous observation that 70% of all VEZF1 sites overlap with GATA2 peaks (Figure 3.13).






Category	Rank	Motif	Logo	E-value
All VEZF1	9th	TTATHW		1.2e-027
VEZF1&Promoter				
VEZF1&Enhancer	2nd	TTATCW		3.8e-024
All GATA2	1st	TTATCW		1.8e-304
VEZF1&GATA2	8th	TTATYW		1.2e-022
VEZF1&Promoter &GATA2				
VEZF1&Enhancer& GATA2	1st	TTATCW		1.9e-029

Table 3.4 GATA motifs enriched at CUT&Tag peaks.

VEZF1 and GATA2 peaks were identified from aligned reads using MACS2. VEZF1 peaks were then categorized as to whether they overlap with GATA2 peaks and/or promoters or enhancers. For each category, MEME-ChIP was used to identify enriched sequence motifs in the 200 bp immediately surrounding the peak summits. The resulting motifs were ranked according to E-value, which is the motif P value times the number of candidate motifs tested. The table shows GATA consensus motifs that were enriched in each category, their rank, logo and E value.

Table 3.5 shows additional motifs enriched in more than one category. These are likely to correspond to

other transcription factors that co-localise with VEZF1. CCAAT box (or ATTGG), the binding motif of nuclear factor Y (NF-Y), is enriched at VEZF1 and GATA2 peaks that are found at promoters, but is not enriched at enhancers (Table 3.6). Motifs for Activator protein 1 (AP-1) and E26 transformation specific (ETS) family members were also enriched in several categories (Table 3.5).

Category	All VEZF1 (Rank, E-value)	VEZF1&Promoter (Rank, E-value)	VEZF1&Enhancer (Rank, E-value)	All GATA2 (Rank, E-value)	VEZF1&GATA2 (Rank, E-value)	VEZF1&Promoter&GATA2 (Rank, E-value)	VEZF1&Enhancer&GATA2 (Rank, E-value)
CCCGCCY/CYCCDCC	1st, 2.2e-296	1st, 7.1e-265	1st, 6.8e-027	2nd, 7.0e-165	1st, 4.3e-229	1st, 3.6e-212	2nd, 1.5e-014
C/CCMCRC/CC/RGGCG	3rd, 4.3e-070	3rd, 6.7e-110		4th, 2.0e-066	3rd, 3.5e-071	3rd, 4.9e-066	
GGR... (SP/KLF family)	6th, 1.3e-038				10th, 2.7e-021	8th, 1.1e-024	
CCAATVR/RCCAATCR/A TTGGYY (NFY, HOX)	2nd, 1.4e-106	2nd, 3.8e-121		5th, 2.1e-038	2nd, 3.8e-103	2nd, 4.7e-113	
ACGTSAY (AP-1 family, CREB family)	7th, 1.8e-035	4th, 1.6e-045				4th, 1.4e-039	
AMGTSAC (ARNTL, ARNT2, MIT/TF E family, AP-1 family)					4th, 1.0e-039		
RTGASTCA (AP-1 family)			5th, 1.4e-007				4th, 8.9e-005
WGATAA/TTATCW/TTA THWTTATYW (GATA family)	9th, 1.2e-027		2nd, 3.8e-024	1st, 1.8e-304	8th, 1.2e-022		1st, 1.9e-029
GGGWGGGR/RGGAGG R (MAZ, SP/KLF family)			4th, 7.9e-008		7th, 1.8e-027		
CAYTCC/CDCYTCC/S HGGAA/VGGAAR/MGG AAR (ETS family)	4th, 2.1e-049 8th, 2.6e-032	5th, 4.6e-042		3rd, 6.0e-083	6th, 4.3e-030	5th, 3.0e-038	
CACAGH (ZIC1/2)			3rd, 3.2e-010				
CBTCCC (SP-5, MAZ, E2F6)							3rd, 2.3e-007
SYGGGA (RBPJ, Thap11, E2F family)		8th, 1.8e-021		9th, 4.9e-029			
Peaks (Regions)	14320	9983	1998	18798	10053	7812	1249

Table 3.5 Summary of motifs enriched at VEZF1 and GATA2 binding sites.

VEZF1 and GATA2 peaks were identified from aligned reads using MACS2. VEZF1 peaks were then categorized as to whether they overlap with GATA2 peaks and/or promoters or enhancers. For each category, MEME-ChIP was used to identify enriched sequence motifs in the 200 bp immediately surrounding the peak summits. The resulting motifs were ranked according to E-value, which is the motif P value times the number of candidate motifs tested. The table shows consensus motifs that were enriched in more than one category, and their rank and E value.






Category	Rank	Motif	Logo	E-value
All VEZF1	2nd	ATTGGYY		1.4e-106
VEZF1&Promoter	2nd	CCAATVR		3.8e-121
VEZF1&Enhancer				
All GATA2	5th	RCCAATCR		2.1e-038
VEZF1&GATA2	2nd	ATTGGYY		3.8e-103
VEZF1&Promoter &GATA2	2nd	CCAATVR		4.7e-113
VEZF1&Enhancer& GATA2				

Table 3.6 CCAAT box-like motifs enriched at CUT&Tag peaks.

VEZF1-like motifs enriched at CUT&Tag peaks VEZF1 and GATA2 peaks were identified from aligned reads using MACS2. VEZF1 peaks were then categorized as to whether they overlap with GATA2 peaks and/or promoters or enhancers. For each category, MEME-ChIP was used to identify enriched sequence motifs in the 200 bp immediately surrounding the peak summits. The resulting motifs were ranked according to E-value, which is the motif P value times the number of candidate motifs tested. The table shows CCAAT-box consensus motifs that were enriched in each category, their rank, logo and E value.

Chapter 4 Discussion

As a novel method evolved from ChIP-Seq, CUT&Tag has significant advantages in terms of cell number requirements, workflow length, and sensitivity. However, there are some differences between CUT&Tag and ChIP-Seq. A major one is that CUT&Tag is performed on live permeabilized cells or isolated nuclei while the starting material for ChIP-Seq is the cells or tissue that are crosslinked with formaldehyde. When applying CUT&Tag to analysis of proteins in nuclei, it is necessary to check whether the antibodies enter nuclei and bind specifically to the targets. In this study, we inspected the applicability of CUT&Tag to mapping VEZF1 binding sites in K562 cells, compared the CUT&Tag data to the ChIP-Seq data from previous research, and explored the interaction between VEZF1, GATA2 and other TFs.

Before performing CUT&Tag, western blotting and immunofluorescence were applied to select antibodies with high activity and specificity. Western blotting is an effective method to determine whether an antibody is active and specific as bands on the blots demonstrate antibodies binding to their targets. However, proteins in the samples of western blotting are denatured, which is different from the status of native proteins in CUT&Tag. Therefore, the binding of antibodies to native proteins was inspected by immunofluorescence. Another reason was that the concentrations of antibodies required for CUT&Tag is reported to be similar to that used in immunofluorescence. When comparing VEZF1 antibodies, AGW-3642-new showed the highest activity and specificity in western blotting, and lower background than NBP1-84301-25ul in immunofluorescence. As for GATA2 antibodies, ab153820 was the only one that showed correct bands in western blotting. Therefore, AGW-3642-new, ab153820, as well as other CTCF and histone modification antibodies were selected for CUT&Tag.

After acquiring the raw data from NGS, the quality was checked and controlled by FastQC. The per base sequence quality decreased while the content of adapters increased towards the 3'end of the reads. Therefore, the adapter sequences were trimmed off to improve the quality of the data. The reads were then aligned to the Hg19 genome by Bowtie2 and converted to bigwig files by bamCoverage.

The alignments were visualized by uploading the bigwig files to the UCSC genome browser as custom tracks. HBA1 and TAL1 locus were chosen to evaluate the quality of bigwig tracks because they are both genes expressed during haematopoiesis and/or vasculogenesis, and the binding of VEZF1 at these loci

has been studied previously. HBA1 encodes alpha-globin, a subunit of hemoglobin. TAL1 is a basic helix-loop-helix transcription factor which is necessary for specification and maturation stages of the three hematopoietic waves. The absence of TAL1 results in the failure in generation of red cells, myeloid cells, megakaryocytes, mast cells, and lymphoid cells (Porcher et al., 1996). TAL1 is also required for the development and remodeling of the vascular network (Porcher et al., 1996; Porcher et al., 2017). All the individual bigwig tracks showed distinct peaks with low background except CnT_H3K27ac_2, which was subsequently discarded. In order to reduce the loss of the peaks that were not identified in all the replicates, bigwig files from replicates were merged prior to peak calling. Compared to the bigwig tracks generated by ChIP-Seq from previous research, the peaks of CUT&Tag tracks were higher, consistent with the lower background signal for of CUT&Tag. However, bigwig tracks of the two methods did not show any difference in the widths of the peaks, which was unexpected.

A high fraction of duplicate reads was discovered in most of the datasets after the main analysis had been completed. To reduce the impact of this issue, it is recommended that larger initial sample sizes are used, with a lower sequencing depth. Duplicate reads should be removed from each bam file before further analysis, and equal numbers of reads from each replicate should be used when merging data from replicates. The removal of duplicates does not greatly change the overall profile of VEZF1 binding, but could affect the inclusion of smaller peaks in the files used for motif analysis. An additional step that would improve the analysis of replicate samples would be to add a small amount of tracer DNA from *E. coli*. This enables normalisation of read counts to the tracer and would remove bias during the merging of data from replicates (Kaya-Okur et al., 2020).

MACS2 was then applied on the merged bigwig files to identify the peaks with a minimum FDR of 0.01 to reduce false positive peaks. The VEZF1 peaks were then overlapped with GATA2 peaks, promoters or enhancers defined by ChromHMM data. More than 64% of VEZF1 peaks intersected with GATA2 peaks at both promoters and enhancers. This is consistent with the analysis of Low (2013), who demonstrated that VEZF1 peaks in K562 cells often co-localised with GATA peaks identified by ENCODE data. Specifically, Low focused on enhancers bound by the TAL1-erythroid complex (TEC) comprising GATA1, TAL1 and NFE2, and showed that 25% of VEZF1-associated enhancers also bound the TEC (Low, 2013, Wadman et al., 1997). Further analysis is needed to investigate whether the GATA2 enhancers identified here also bind other members of the TEC.

The chromatin features of VEZF1 peaks associated with promoter or enhancer regions were analyzed by comparing the enrichment of different histone modifications using the ChIP-Cor tool. Analysis of the promoters showed that H3K4me3, the mark of active promoters, was much more enriched than H3K4me1. This indicated that most genes with VEZF1 bound at their promoters were in active state. The situation is different at enhancers. H3K4me1 is a well-known mark of strong enhancers, but H3K4me1 was also less enriched than H3K4me3 at enhancers according to the graph (Ernst et al., 2011). H3K27ac, which is found at both active promoters and strong enhancers, was also more enriched at promoters compared to enhancers in our data (Ernst et al., 2011). It could be inferred that enhancers bound by VEZF1 and GATA2 have weaker enrichment of H3K4me1 and H3K27ac than expected. However, additional analyses of chromatin data at all promoters and enhancers, and comparison with other histone modification data sets, is required before any conclusions can be drawn. An important caveat is that ChromHMM predicts chromatin state through analysis of nine chromatin marks. It does not include RNA transcript data, DNase I hypersensitive site mapping, or data from other factors that are often bound at enhancers (Ernst and Kellis, 2010; Ernst et al., 2011), and so the prediction of promoters and enhancers may differ from those identified through other methods. In addition, differences between the K562 cells used in this study and the K562 cells used by the Bernstein lab in the ENCODE ChromHMM study could result in variation in gene expression and epigenetic profiles (Ernst et al., 2011).

DNA sequences within ± 100 bp of VEZF1 peak summits were analyzed using MEME-ChIP to find motifs and further explore interactions between VEZF1 and other TFs. The most frequently occurring motif in VEZF1 peaks followed the GGGNGGGG pattern, and this was highly enriched at both promoter- and enhancer-associated VEZF1 peaks, and at those where GATA2 was also bound.

The GGGNGGGG pattern is also recognized by the SP/Krüppel-like factor (KLF) family, a set of TFs that bind to both promoter and enhancer regions of the genome and regulate gene expression by recruiting other regulatory proteins (Fernandez-Zapico et al., 2011). Sp/KLF family consists of 2 subfamilies which are the Sp subfamily, comprising 9 members, and the KLF subfamily, which has 17 members. Although all the Sp/KLF factors recognize CACCC boxes and GC-rich elements, different members have different functions as transcriptional activators or repressors (Funnell et al., 2007). Another TF that recognizes the GGGNGGGG pattern is Myc-associated zinc finger (MAZ) protein. MAZ is essential for erythropoiesis, and

binds to the promoter of the erythroid-specific human α -globin gene. MAZ is also reported to colocalize with GATA1 at enhancer elements, indicating a connection of function between them (Deen et al., 2021). Competition between VEZF1 and other factors such as SP1 and MAZ at individual binding sites is likely to play an important role in the regulation of haematopoiesis and vasculogenesis. Detailed analysis of their binding profiles, performed in the same cells, will be needed to establish how these factors contribute to gene regulation, and a rapid, robust and economical approach such as CUT&Tag will greatly facilitate this type of analysis.

Notably, the VEZF1 consensus of 9(dG.dC) was not identified through motif discovery, which is different from the results of the previous study (Low, 2013). Low found that 9(dG.dC) motifs were found to be highly enriched at VEZF1-associated promoters while GGGGA/TGGGG motifs occurred with greatest frequency at VEZF1-associated enhancers (Low, 2013). These differences between the motifs enriched at VEZF1-associated promoters and enhancers were not discovered in our analyses. Moreover, the co-localisation of VEZF1 with GATA2 did not appear to alter the preferred DNA sequence motif of VEZF1. A potential reason for why the 9(dG.dC) motif was not identified is the difference between the tools used in our study and previous study by Low (2013). For example, the tool we used for motif discovery was MEME-ChIP, and the algorithm used by the tool was DREME. The motifs that DREME searches for are limited to be no wider than 8 bp in order to achieve higher speed, which causes it to miss any longer motifs (Bailey, 2011). In previous study, MEME, which searches for enriched DNA sequence motifs of up to 30 bp, and POSMO, which searches for motifs of 7 to 9 bp were used (Low, 2013). Further investigation could compare the use of MEME-ChIP, MEME, POSMO and other motif identification tools to determine whether any of these offer additional sensitivity. An additional challenge is that VEZF1 often appears to bind at several sites within individual promoters, as peaks can be quite broad. Identifying these individual binding sites could refine the motif analysis and increase the sensitivity for detecting different classes of VEZF1 motif. This could be achieved by modifying the peak detection process using a peak splitting tool. This would split broad regions of enrichment into subpeaks that could then be processed separately. A further refinement would be to categorise peaks on the basis of their enrichment levels in order to determine whether “stronger” binding sites for VEZF1 have a different motif compared to “weaker” ones. Another possibility is that VEZF1 has a higher affinity for the GGGNGGGG motif than 9(dG.dC) *in vivo* due to the influence of other transcription factors and chromatin states. We currently lack solid evidence to reach an explanation for the differences, but this

would be an interesting direction for the research in the future.

Enrichment of other motifs at the VEZF1 peaks provides information about other transcription factors that may work alongside VEZF1 to regulate gene expression, or that may interact with VEZF at these chromatin sites. For example, DNA binding motifs for the activator protein 1 (AP-1) family of transcription factors were enriched in all the categories except the “all GATA2” grouping. AP-1 proteins regulate gene expression typically by binding DNA as homodimers and heterodimers. AP-1 family composed of four subfamilies, which are Jun (JunB, JunD), Fos (FRA-1, FRA-2, and FosB), ATF, and Maf subfamilies (Wu et al., 2021). The motifs we discovered such as ACGTSAY and RTGASTCA are recognized by many members of AP-1 family.

Some transcription factor binding motifs were not present in all the categories of sites analysed. Instead, they were only enriched at either VEZF1- or GATA2-associated promoters or enhancers. Motifs for NF-Y/Hox and ETS factors were enriched at VEZF1-associated promoters but not at the enhancers. NF-Y is a transcription factor that mainly binds at promoter regions and regulates cell cycle progression in proliferating cells (Oldfield et al., 2014). The ETS family includes 28 transcription factors (Wang et al., 2023). They share a conserved winged helix-turn-helix topology which recognizes the 5'-GGA(A/T)-3' motifs. ETS factors regulate a variety of biological processes such as angiogenesis and hematopoiesis by acting as transcriptional activators or repressors (Wang et al., 2023).

At VEZF1-associated enhancers, motifs for ZIC1/2 were also discovered. ZIC is the abbreviation of “zinc finger protein of the cerebellum”, which means that these proteins are mainly expressed in cerebellum and are essential for the development of the cerebellum (Aruga and Millen, 2018). As expected, GATA motifs were enriched at enhancers where GATA2 was bound, but GATA motifs were also enriched in the “all VEZF1 peaks” category and at VEZF1-associated enhancers. Both GATA2 and VEZF1 are important TFs in erythroid regulation. The enrichment of GATA motifs at VEZF1-associated enhancers, and the high proportion of VEZF1 peaks co-localised with GATA2 (64%) discovered in our study provides new evidence for the interaction between VEZF1 and GATA2 at enhancers.

Co-localisation of other factors at VEZF1 peaks was also investigated by Low (2013), who intersected VEZF1 peaks with ENCODE ChIP-seq data for transcription factors in K562 cells. That analysis supports

co-localisation of SP1, ETS1 and AP-1 family members at VEZF-1 associated promoters, and co-localisation of SP1, AP-1 family members, and GATA1 and GATA2 at VEZF1-associated enhancers (Low, 2013). Taken together, the data support the conclusion that VEZF1 works alongside factors such as SP1, GATA1/2, AP-1 family members and ETS family member to regulate the same genes in K562 cells. Further investigation is needed to see whether there are physical interactions between these factors, and how they influence the binding of each other to chromatin sites *in vivo*. For example, co-immunoprecipitation of VEZF1 complexes followed by SILAC, a quantitative mass spectrometry approach, could be used to identify VEZF1 interaction partners. Genome-wide analyses of VEZF1 binding in cells where GATA1/2 or MAZ have been knocked out using CRISPR, and complimentary analyses of GATA1//2/MAZ binding in *VeZF1* knockout cells, are needed to determine how these factors influence each others' binding *in vivo*, and this study demonstrates that CUT&Tag would be a valuable tool for these investigations.

In summary, this study shows that CUT&Tag is a more efficient method for mapping the binding sites of transcription factors and histone modifications compared to ChIP-seq, and that it generates data that is suitable for subsequent peak and motif analysis. CUT&Tag shows significant advantages in terms of the amount of cells required, the cost, the time of the procedure, and the quality of the data. Although CUT&Tag requires higher specificity of antibodies, and may produce lower resolution than CUT&RUN when mapping TFs due to the larger size of pA-Tn5 complex than MNase and requirement for more stringent washes (Kaya-Okur et al., 2020), this rapid and economical approach can now be used for in depth, systematic analysis of the role of VEZF1 and other factors in regulating gene expression during the differentiation of endothelial cells in vasculogenesis.

References

- Aitsebaomo J, Kingsley-Kallesen ML, Wu Y, Quertermous T, Patterson C. Vezf1/DB1 is an endothelial cell-specific transcription factor that regulates expression of the endothelin-1 promoter. *J Biol Chem*. 2001 Oct 19;276(42):39197-205. doi: 10.1074/jbc.M105166200. Epub 2001 Aug 14. PMID: 11504723.
- AlAbdi L, He M, Yang Q, Norvil AB, Gowher H. The transcription factor Vezf1 represses the expression of the antiangiogenic factor Cited2 in endothelial cells. *J Biol Chem*. 2018 Jul 13;293(28):11109-11118. doi: 10.1074/jbc.RA118.002911. Epub 2018 May 24. PMID: 29794136; PMCID: PMC6052231.
- Al-Hosni A. Roles for the VEZF1 transcription factor in erythroid and housekeeping gene expression. 2016.
- Aliya M R Al-Hosni. Roles for the VEZF1 transcription factor in erythroid and housekeeping gene expression. 2016.
- Ambrosini G, Dreos R, Kumar S, Bucher P. The ChIP-Seq tools and web server: a resource for analyzing ChIP-Seq and other types of genomic data. *BMC Genomics*. 2016 Nov 18;17(1):938. doi: 10.1186/s12864-016-3288-8. PMID: 27863463; PMCID: PMC5116162.
- Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010 Apr 26. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Aruga J, Millen KJ. ZIC1 Function in Normal Cerebellar Development and Human Developmental Pathology. *Adv Exp Med Biol*. 2018;1046:249-268. doi: 10.1007/978-981-10-7311-3_13. PMID: 29442326.
- Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*. 2011 Jun 15;27(12):1653-9. doi: 10.1093/bioinformatics/btr261. Epub 2011 May 4. PMID: 21543442; PMCID: PMC3106199.
- Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. *Nucleic Acids Res*. 2015 Jul 1;43(W1):W39-49. doi: 10.1093/nar/gkv416. Epub 2015 May 7. PMID: 25953851; PMCID: PMC4489269.
- Bell AC, Felsenfeld G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature*. 2000 May 25;405(6785):482-5. doi: 10.1038/35013100. PMID: 10839546.
- Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*. 1999 Aug 6;98(3):387-96. doi: 10.1016/s0092-8674(00)81967-4. PMID: 10458613.
- Broad Institute of MIT and Harvard. Picard. In Broad Institute, GitHub repository. GitHub. 2014 Sep 9. <http://broadinstitute.github.io/picard/>
- Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? *Mol Cell*. 2013 Mar 7;49(5):825-37. doi: 10.1016/j.molcel.2013.01.038. PMID: 23473601; PMCID: PMC3857148.
- Calvo KR, Hickstein DD. The spectrum of GATA2 deficiency syndrome. *Blood*. 2023 Mar 30;141(13):1524-1532. doi: 10.1182/blood.2022017764. PMID: 36455197; PMCID: PMC10082373.
- Chan WY, Follows GA, Lacaud G, Pimanda JE, Landry JR, Kinston S, Knezevic K, Piltz S, Donaldson IJ, Gambardella L, Sablitzky F, Green AR, Kouskoff V, Göttgens B. The paralogous hematopoietic regulators *Lyl1* and *Scl* are coregulated by *Ets* and *GATA* factors, but *Lyl1* cannot rescue the early *Scl*^{-/-} phenotype. *Blood*. 2007 Mar 1;109(5):1908-16. doi: 10.1182/blood-2006-05-023226. Epub 2006 Oct 19. PMID: 17053063.
- Choi K. The hemangioblast: a common progenitor of hematopoietic and endothelial cells. *J Hematother Stem Cell Res*. 2002 Feb;11(1):91-101. doi: 10.1089/152581602753448568. PMID: 11847006.
- Deans C, Maggert KA. What do you mean, "epigenetic"? *Genetics*. 2015 Apr;199(4):887-96. doi: 10.1534/genetics.114.173492. PMID: 25855649; PMCID: PMC4391566.

- Deen D, Butter F, Daniels DE, Ferrer-Vicens I, Ferguson DCJ, Holland ML, Samara V, Sloane-Stanley JA, Ayyub H, Mann M, Frayne J, Garrick D, Vernimmen D. Identification of the transcription factor MAZ as a regulator of erythropoiesis. *Blood Adv.* 2021 Aug 10;5(15):3002-3015. doi: 10.1182/bloodadvances.2021004609. Erratum in: *Blood Adv.* 2022 Mar 22;6(6):1854. PMID: 34351390; PMCID: PMC8361462.
- Dehingia B, Milewska M, Janowski M, Pękowska A. CTCF shapes chromatin structure and gene expression in health and disease. *EMBO Rep.* 2022 Sep 5;23(9):e55146. doi: 10.15252/embr.202255146. Epub 2022 Aug 22. PMID: 35993175; PMCID: PMC9442299.
- Dickson J, Gowher H, Strogantsev R, Gaszner M, Hair A, Felsenfeld G, West AG. VEZF1 elements mediate protection from DNA methylation. *PLoS Genet.* 2010 Jan;6(1):e1000804. doi: 10.1371/journal.pgen.1000804. Epub 2010 Jan 8. PMID: 20062523; PMCID: PMC2795164.
- Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature.* 2012 Apr 11;485(7398):376-80. doi: 10.1038/nature11082. PMID: 22495300; PMCID: PMC3356448.
- Dobrzycki T, Lalwani M, Telfer C, Monteiro R, Patient R. The roles and controls of GATA factors in blood and cardiac development. *IUBMB Life.* 2020 Jan;72(1):39-44. doi: 10.1002/iub.2178. Epub 2019 Nov 28. PMID: 31778014; PMCID: PMC6973044.
- Du K, Zhao C, Wang L, Wang Y, Zhang KZ, Shen XY, Sun HX, Gao W, Lu X. MiR-191 inhibit angiogenesis after acute ischemic stroke targeting VEZF1. *Aging (Albany NY).* 2019 May 7;11(9):2762-2786. doi: 10.18632/aging.101948. PMID: 31064890; PMCID: PMC6535071.
- El-Kady A, Klenova E. Regulation of the transcription factor, CTCF, by phosphorylation with protein kinase CK2. *FEBS Lett.* 2005 Feb 28;579(6):1424-34. doi: 10.1016/j.febslet.2005.01.044. PMID: 15733852.
- El Omari K, Hoosdally SJ, Tuladhar K, Karia D, Hall-Ponsel e E, Platonova O, Vyas P, Patient R, Porcher C, Mancini EJ. Structural basis for LMO2-driven recruitment of the SCL:E47bHLH heterodimer to hematopoietic-specific transcriptional targets. *Cell Rep.* 2013 Jul 11;4(1):135-47. doi: 10.1016/j.celrep.2013.06.008. Epub 2013 Jul 3. PMID: 23831025; PMCID: PMC3714592.
- Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol.* 2010 Aug;28(8):817-25. doi: 10.1038/nbt.1662. Epub 2010 Jul 25. PMID: 20657582; PMCID: PMC2919626.
- Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature.* 2011 May 5;473(7345):43-9. doi: 10.1038/nature09906. Epub 2011 Mar 23. PMID: 21441907; PMCID: PMC3088773.
- Evans T, Reitman M, Felsenfeld G. An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes. *Proc Natl Acad Sci U S A.* 1988 Aug;85(16):5976-80. doi: 10.1073/pnas.85.16.5976. PMID: 3413070; PMCID: PMC281888.
- Fernandez-Zapico ME, Lomberk GA, Tsuji S, DeMars CJ, Bardsley MR, Lin YH, Almada LL, Han JJ, Mukhopadhyay D, Ordog T, Buttar NS, Urrutia R. A functional family-wide screening of SP/KLF proteins identifies a subset of suppressors of KRAS-mediated cell growth. *Biochem J.* 2011 Apr 15;435(2):529-37. doi: 10.1042/BJ20100773. PMID: 21171965; PMCID: PMC3130109.
- Funnell AP, Maloney CA, Thompson LJ, Keys J, Tallack M, Perkins AC, Crossley M. Erythroid Kr uppel-like factor directly activates the basic Kr uppel-like factor gene in erythroid cells. *Mol Cell Biol.* 2007 Apr;27(7):2777-90. doi: 10.1128/MCB.01658-06. Epub 2007 Feb 5. PMID: 17283065; PMCID: PMC1899893.
- Gassler J, Brand o HB, Imakaev M, Flyamer IM, Ladst atter S, Bickmore WA, Peters JM, Mirny LA, Tachibana K. A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J.* 2017 Dec 15;36(24):3600-3618. doi: 10.15252/embj.201798083. Epub 2017

- Dec 7. PMID: 29217590; PMCID: PMC5730859.
- Gerald D, Adini I, Shechter S, Perruzzi C, Varnau J, Hopkins B, Kazerounian S, Kurschat P, Blachon S, Khedkar S, Bagchi M, Sherris D, Prendergast GC, Klagsbrun M, Stuhlmann H, Rigby AC, Nagy JA, Benjamin LE. RhoB controls coordination of adult angiogenesis and lymphangiogenesis following injury by regulating VEZF1-mediated transcription. *Nat Commun.* 2013;4:2824. doi: 10.1038/ncomms3824. PMID: 24280686; PMCID: PMC3868161.
- Goldie LC, Nix MK, Hirschi KK. Embryonic vasculogenesis and hematopoietic specification. *Organogenesis.* 2008 Oct;4(4):257-63. doi: 10.4161/org.4.4.7416. PMID: 19337406; PMCID: PMC2634331.
- Gordon DF, Lewis SR, Haugen BR, James RA, McDermott MT, Wood WM, Ridgway EC. Pit-1 and GATA-2 interact and functionally cooperate to activate the thyrotropin beta-subunit promoter. *J Biol Chem.* 1997 Sep 26;272(39):24339-47. doi: 10.1074/jbc.272.39.24339. PMID: 9305891.
- Gowher H, Stuhlmann H, Felsenfeld G. Vezf1 regulates genomic DNA methylation through its effects on expression of DNA methyltransferase Dnmt3b. *Genes Dev.* 2008 Aug 1;22(15):2075-84. doi: 10.1101/gad.1658408. PMID: 18676812; PMCID: PMC2492749.
- Haering CH, Farcas AM, Arumugam P, Metson J, Nasmyth K. The cohesin ring concatenates sister DNA molecules. *Nature.* 2008 Jul 17;454(7202):297-301. doi: 10.1038/nature07098. Epub 2008 Jul 2. PMID: 18596691.
- Haarhuis JHI, van der Weide RH, Blomen VA, Yáñez-Cuna JO, Amendola M, van Ruiten MS, Krijger PHL, Teunissen H, Medema RH, van Steensel B, Brummelkamp TR, de Wit E, Rowland BD. The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell.* 2017 May 4;169(4):693-707.e14. doi: 10.1016/j.cell.2017.04.013. PMID: 28475897; PMCID: PMC5422210.
- Hansen AS, Hsieh TS, Cattoglio C, Pustova I, Saldaña-Meyer R, Reinberg D, Darzacq X, Tjian R. Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF. *Mol Cell.* 2019 Nov 7;76(3):395-411.e13. doi: 10.1016/j.molcel.2019.07.039. Epub 2019 Sep 12. PMID: 31522987; PMCID: PMC7251926.
- Harvard Chan Bioinformatics Core (HBC)'s teaching team. Introduction to RNA-seq using high-performance computing (HPC). 2021 Mar 23. https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html
- Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol.* 2015 Mar;16(3):144-54. doi: 10.1038/nrm3949. Epub 2015 Feb 4. PMID: 25650801; PMCID: PMC4517609.
- Howe FS, Fischl H, Murray SC, Mellor J. Is H3K4me3 instructive for transcription activation? *Bioessays.* 2017 Jan;39(1):1-12. doi: 10.1002/bies.201600095. Epub 2016 Nov 7. PMID: 28004446.
- Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, Taatjes DJ, Dekker J, Young RA. Mediator and cohesin connect gene expression and chromatin architecture. *Nature.* 2010 Sep 23;467(7314):430-5. doi: 10.1038/nature09380. Epub 2010 Aug 18. Erratum in: *Nature.* 2011 Apr 14;472(7342):247. PMID: 20720539; PMCID: PMC2953795.
- Karnuta JM, Scacheri PC. Enhancers: bridging the gap between gene control and human disease. *Hum Mol Genet.* 2018 Aug 1;27(R2):R219-R227. doi: 10.1093/hmg/ddy167. PMID: 29726898; PMCID: PMC6061867.
- Katsumura KR, Bresnick EH; GATA Factor Mechanisms Group. The GATA factor revolution in hematology. *Blood.* 2017 Apr 13;129(15):2092-2102. doi: 10.1182/blood-2016-09-687871. Epub 2017 Feb 8. PMID: 28179282; PMCID: PMC5391619.
- Kaya-Okur HS, Janssens DH, Henikoff JG, Ahmad K, Henikoff S. Efficient low-cost chromatin profiling with CUT&Tag. *Nat Protoc.* 2020 Oct;15(10):3264-3283. doi: 10.1038/s41596-020-0373-x. Epub 2020 Sep 10. PMID: 32913232; PMCID: PMC8318778.

- Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, Ahmad K, Henikoff S. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun.* 2019 Apr 29;10(1):1930. doi: 10.1038/s41467-019-09982-5. PMID: 31036827; PMCID: PMC6488672.
- Khan A, Zhang X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.* 2016 Jan 4;44(D1):D164-71. doi: 10.1093/nar/gkv1002. Epub 2015 Oct 4. PMID: 26438538; PMCID: PMC4702767.
- Khandekar M, Brandt W, Zhou Y, Dagenais S, Glover TW, Suzuki N, Shimizu R, Yamamoto M, Lim KC, Engel JD. A Gata2 intronic enhancer confers its pan-endothelia-specific regulation. *Development.* 2007 May;134(9):1703-12. doi: 10.1242/dev.001297. Epub 2007 Mar 29. PMID: 17395646.
- Koga S, Yamaguchi N, Abe T, Minegishi M, Tsuchiya S, Yamamoto M, Minegishi N. Cell-cycle-dependent oscillation of GATA2 expression in hematopoietic cells. *Blood.* 2007 May 15;109(10):4200-8. doi: 10.1182/blood-2006-08-044149. Epub 2007 Jan 25. PMID: 17255359.
- Ko LJ, Engel JD. DNA-binding specificities of the GATA transcription factor family. *Mol Cell Biol.* 1993 Jul;13(7):4011-22. doi: 10.1128/mcb.13.7.4011-4022.1993. PMID: 8321208; PMCID: PMC359950.
- Koyano-Nakagawa N, Nishida J, Baldwin D, Arai K, Yokota T. Molecular cloning of a novel human cDNA encoding a zinc finger protein that binds to the interleukin-3 promoter. *Mol Cell Biol.* 1994 Aug;14(8):5099-107. doi: 10.1128/mcb.14.8.5099-5107.1994. PMID: 8035792; PMCID: PMC359028.
- Koyunlar C, Gioacchino E, Vadgama D, de Looper H, Zink J, Ter Borg MND, Hoogenboezem R, Havermans M, Sanders MA, Bindels E, Dzierzak E, Touw IP, de Pater E. Gata2-regulated Gfi1b expression controls endothelial programming during endothelial-to-hematopoietic transition. *Blood Adv.* 2023 May 23;7(10):2082-2093. doi: 10.1182/bloodadvances.2022008019. PMID: 36649572; PMCID: PMC10196789.
- Kuhnert F, Campagnolo L, Xiong JW, Lemons D, Fitch MJ, Zou Z, Kiosses WB, Gardner H, Stuhlmann H. Dosage-dependent requirement for mouse *Vezf1* in vascular system development. *Dev Biol.* 2005 Jul 1;283(1):140-56. doi: 10.1016/j.ydbio.2005.04.003. PMID: 15882861; PMCID: PMC1453095.
- Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. The Human Transcription Factors. *Cell.* 2018 Feb 8;172(4):650-665. doi: 10.1016/j.cell.2018.01.029. Erratum in: *Cell.* 2018 Oct 4;175(2):598-599. PMID: 29425488.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923. PMID: 22388286; PMCID: PMC3322381.
- Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):R25. doi: 10.1186/gb-2009-10-3-r25. Epub 2009 Mar 4. PMID: 19261174; PMCID: PMC2690996.
- Lefevre P, Witham J, Lacroix CE, Cockerill PN, Bonifer C. The LPS-induced transcriptional upregulation of the chicken lysozyme locus involves CTCF eviction and noncoding RNA transcription. *Mol Cell.* 2008 Oct 10;32(1):129-39. doi: 10.1016/j.molcel.2008.07.023. PMID: 18851839; PMCID: PMC2581490.
- Lewis CD, Clark SP, Felsenfeld G, Gould H. An erythrocyte-specific protein that binds to the poly(dG) region of the chicken beta-globin gene promoter. *Genes Dev.* 1988 Jul;2(7):863-73. doi: 10.1101/gad.2.7.863. PMID: 3209071.
- Li Y, Haarhuis JHI, Sedeño Cacciatore Á, Oldenkamp R, van Ruiten MS, Willems L, Teunissen H, Muir KW, de Wit E, Rowland BD, Panne D. The structural basis for cohesin-CTCF-anchored loops. *Nature.* 2020 Feb;578(7795):472-476. doi: 10.1038/s41586-019-1910-z. Epub 2020 Jan 6. PMID: 31905366; PMCID: PMC7035113.
- Lobanenkov VV, Nicolas RH, Adler VV, Paterson H, Klenova EM, Polotskaja AV, Goodwin GH. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene.* 1990 Dec;5(12):1743-53. PMID: 2284094.

- Low C. Genomic Interactions of the Transcription Factor VEZF1. PhD thesis. University of Glasgow. 2013. <https://theses.gla.ac.uk/5078/>
- Martinez SR, Miranda JL. CTCF terminal segments are unstructured. *Protein Sci.* 2010 May;19(5):1110-6. doi: 10.1002/pro.367. PMID: 20196073; PMCID: PMC2868253.
- Maurya SS. Role of Enhancers in Development and Diseases. *Epigenomes.* 2021 Oct 4;5(4):21. doi: 10.3390/epigenomes5040021. PMID: 34968246; PMCID: PMC8715447.
- Ma W, Noble WS, Bailey TL. Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nat Protoc.* 2014;9(6):1428-50. doi: 10.1038/nprot.2014.083. Epub 2014 May 22. PMID: 24853928; PMCID: PMC4175909.
- Meers MP, Tenenbaum D, Henikoff S. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. *Epigenetics Chromatin.* 2019 Jul 12;12(1):42. doi: 10.1186/s13072-019-0287-4. PMID: 31300027; PMCID: PMC6624997.
- Miyashita H, Kanemura M, Yamazaki T, Abe M, Sato Y. Vascular endothelial zinc finger 1 is involved in the regulation of angiogenesis: possible contribution of stathmin/OP18 as a downstream target gene. *Arterioscler Thromb Vasc Biol.* 2004 May;24(5):878-84. doi: 10.1161/01.ATV.0000126373.52450.32. Epub 2004 Mar 18. PMID: 15031128.
- Oldfield AJ, Yang P, Conway AE, Cinghu S, Freudenberg JM, Yellaboina S, Jothi R. Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors. *Mol Cell.* 2014 Sep 4;55(5):708-22. doi: 10.1016/j.molcel.2014.07.005. Epub 2014 Aug 14. PMID: 25132174; PMCID: PMC4157648.
- Orkin SH. Transcription factors and hematopoietic development. *J Biol Chem.* 1995 Mar 10;270(10):4955-8. doi: 10.1074/jbc.270.10.4955. PMID: 7890597.
- Park C, Kim TM, Malik AB. Transcriptional regulation of endothelial cell and vascular development. *Circ Res.* 2013 May 10;112(10):1380-400. doi: 10.1161/CIRCRESAHA.113.301078. PMID: 23661712; PMCID: PMC3730491.
- Patan S. Vasculogenesis and angiogenesis. *Cancer Treat Res.* 2004;117:3-32. doi: 10.1007/978-1-4419-8871-3_1. PMID: 15015550.
- Peixoto P, Cartron PF, Serandour AA, Hervouet E. From 1957 to Nowadays: A Brief History of Epigenetics. *Int J Mol Sci.* 2020 Oct 14;21(20):7571. doi: 10.3390/ijms21207571. PMID: 33066397; PMCID: PMC7588895.
- Porcher C, Chagraoui H, Kristiansen MS. SCL/TAL1: a multifaceted regulator from blood development to disease. *Blood.* 2017 Apr 13;129(15):2051-2060. doi: 10.1182/blood-2016-12-754051. Epub 2017 Feb 8. PMID: 28179281.
- Porcher C, Swat W, Rockwell K, Fujiwara Y, Alt FW, Orkin SH. The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell.* 1996 Jul 12;86(1):47-57. doi: 10.1016/s0092-8674(00)80076-8. PMID: 8689686.
- Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 2016 Jul 8;44(W1):W160-5. doi: 10.1093/nar/gkw257. Epub 2016 Apr 13. PMID: 27079975; PMCID: PMC4987876.
- Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell.* 2014 Dec 18;159(7):1665-80. doi: 10.1016/j.cell.2014.11.021. Epub 2014 Dec 11. Erratum in: *Cell.* 2015 Jul 30;162(3):687-8. PMID: 25497547; PMCID: PMC5635824.
- Recillas-Targa F, Pikaart MJ, Burgess-Beusse B, Bell AC, Litt MD, West AG, Gaszner M, Felsenfeld G. Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proc Natl Acad Sci U S A.* 2002 May 14;99(10):6883-8. doi: 10.1073/pnas.102179399.

PMID: 12011446; PMCID: PMC124498.

- Renda M, Baglivo I, Burgess-Beusse B, Esposito S, Fattorusso R, Felsenfeld G, Pedone PV. Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. *J Biol Chem*. 2007 Nov 16;282(46):33336-33345. doi: 10.1074/jbc.M706213200. Epub 2007 Sep 7. PMID: 17827499.
- Rivera Gonzalez. Investigating the role of the human transcription factor VEZF1 in erythroid and vascular endothelial differentiation, 2017.
- Rodríguez-Ubrega J, Ballestar E. Chromatin immunoprecipitation. *Methods Mol Biol*. 2014;1094:309-18. doi: 10.1007/978-1-62703-706-8_24. PMID: 24162998.
- Ruslan S. Strogantsev. Mapping and characterisation of genomic binding sites of the chromatin barrier protein VEZF1. 2009.
- Saldaña-Meyer R, González-Buendía E, Guerrero G, Narendra V, Bonasio R, Recillas-Targa F, Reinberg D. CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. *Genes Dev*. 2014 Apr 1;28(7):723-34. doi: 10.1101/gad.236869.113. PMID: 24696455; PMCID: PMC4015496.
- Saldaña-Meyer R, Rodríguez-Hernaez J, Escobar T, Nishana M, Jácome-López K, Nora EP, Bruneau BG, Tsirigos A, Furlan-Magaril M, Skok J, Reinberg D. RNA Interactions Are Essential for CTCF-Mediated Genome Organization. *Mol Cell*. 2019 Nov 7;76(3):412-422.e5. doi: 10.1016/j.molcel.2019.08.015. Epub 2019 Sep 12. PMID: 31522988; PMCID: PMC7195841.
- Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NC, Schreiber SL, Mellor J, Kouzarides T. Active genes are tri-methylated at K4 of histone H3. *Nature*. 2002 Sep 26;419(6905):407-11. doi: 10.1038/nature01080. Epub 2002 Sep 11. PMID: 12353038.
- Sharifi-Zarchi A, Gerovska D, Adachi K, Totonchi M, Pezeshk H, Taft RJ, Schöler HR, Chitsaz H, Sadeghi M, Baharvand H, Araúzo-Bravo MJ. DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. *BMC Genomics*. 2017 Dec 12;18(1):964. doi: 10.1186/s12864-017-4353-7. PMID: 29233090; PMCID: PMC5727985.
- Shimamoto T, Ohyashiki K, Ohyashiki JH, Kawakubo K, Fujimura T, Iwama H, Nakazawa S, Toyama K. The expression pattern of erythrocyte/megakaryocyte-related transcription factors GATA-1 and the stem cell leukemia gene correlates with hematopoietic differentiation and is associated with outcome of acute myeloid leukemia. *Blood*. 1995 Oct 15;86(8):3173-80. PMID: 7579412.
- Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schübeler D. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*. 2011 Dec 14;480(7378):490-5. doi: 10.1038/nature10716. Erratum in: *Nature*. 2012 Apr 26;484(7395):550. van Nimwegen, Erik [added]. PMID: 22170606.
- Steger DJ, Hecht JH, Mellon PL. GATA-binding proteins regulate the human gonadotropin alpha-subunit gene in the placenta and pituitary gland. *Mol Cell Biol*. 1994 Aug;14(8):5592-602. doi: 10.1128/mcb.14.8.5592-5602.1994. PMID: 7518566; PMCID: PMC359078.
- Suárez Y, Fernández-Hernando C, Yu J, Gerber SA, Harrison KD, Pober JS, Iruela-Arispe ML, Merckenschlager M, Sessa WC. Dicer-dependent endothelial microRNAs are necessary for postnatal angiogenesis. *Proc Natl Acad Sci U S A*. 2008 Sep 16;105(37):14082-7. doi: 10.1073/pnas.0804597105. Epub 2008 Sep 8. PMID: 18779589; PMCID: PMC2544582.
- Sun S, Del Rosario BC, Szanto A, Ogawa Y, Jeon Y, Lee JT. Jpx RNA activates Xist by evicting CTCF. *Cell*. 2013 Jun 20;153(7):1537-51. doi: 10.1016/j.cell.2013.05.028. PMID: 23791181; PMCID: PMC3777401.
- Tam PP, Behringer RR. Mouse gastrulation: the formation of a mammalian body plan. *Mech Dev*. 1997 Nov;68(1-2):3-25. doi: 10.1016/s0925-4773(97)00123-8. PMID: 9431800.
- Tsai FY, Keller G, Kuo FC, Weiss M, Chen J, Rosenblatt M, Alt FW, Orkin SH. An early haematopoietic

- defect in mice lacking the transcription factor GATA-2. *Nature*. 1994 Sep 15;371(6494):221-6. doi: 10.1038/371221a0. PMID: 8078582.
- Tsai SF, Martin DI, Zon LI, D'Andrea AD, Wong GG, Orkin SH. Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. *Nature*. 1989 Jun 8;339(6224):446-51. doi: 10.1038/339446a0. PMID: 2725678.
- Wadman IA, Osada H, Grütz GG, Agulnick AD, Westphal H, Forster A, Rabbitts TH. The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J*. 1997 Jun 2;16(11):3145-57. doi: 10.1093/emboj/16.11.3145. PMID: 9214632; PMCID: PMC1169933.
- Wang Y, Huang Z, Sun M, Huang W, Xia L. ETS transcription factors: Multifaceted players from cancer progression to tumor immunity. *Biochim Biophys Acta Rev Cancer*. 2023 May;1878(3):188872. doi: 10.1016/j.bbcan.2023.188872. Epub 2023 Feb 24. PMID: 36841365.
- Wen Z, Zhang L, Ruan H, Li G. Histone variant H2A.Z regulates nucleosome unwrapping and CTCF binding in mouse ES cells. *Nucleic Acids Res*. 2020 Jun 19;48(11):5939-5952. doi: 10.1093/nar/gkaa360. PMID: 32392318; PMCID: PMC7293034.
- Whitcomb J, Gharibeh L, Nemer M. From embryogenesis to adulthood: Critical role for GATA factors in heart development and function. *IUBMB Life*. 2020 Jan;72(1):53-67. doi: 10.1002/iub.2163. Epub 2019 Sep 13. PMID: 31520462.
- Wozniak RJ, Keles S, Lugus JJ, Young KH, Boyer ME, Tran TM, Choi K, Bresnick EH. Molecular hallmarks of endogenous chromatin complexes containing master regulators of hematopoiesis. *Mol Cell Biol*. 2008 Nov;28(21):6681-94. doi: 10.1128/MCB.01061-08. Epub 2008 Sep 8. PMID: 18779319; PMCID: PMC2573226.
- Wu Ct, Morris JR. Genes, genetics, and epigenetics: a correspondence. *Science*. 2001 Aug 10;293(5532):1103-5. doi: 10.1126/science.293.5532.1103. PMID: 11498582.
- Wu H, Luo YX, Hu W, Zhao ML, Bie J, Yang M, Pan R, Huang NX, Feng G, Liu K, Song G. MicroRNA-382-5p inhibits osteosarcoma development and progression by negatively regulating VEZF1 expression. *Oncol Lett*. 2021 Nov;22(5):752. doi: 10.3892/ol.2021.13013. Epub 2021 Aug 27. PMID: 34539856; PMCID: PMC8436354.
- Wutz G, Várnai C, Nagasaka K, Cisneros DA, Stocsits RR, Tang W, Schoenfelder S, Jessberger G, Muhar M, Hossain MJ, Walther N, Koch B, Kueblbeck M, Ellenberg J, Zuber J, Fraser P, Peters JM. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J*. 2017 Dec 15;36(24):3573-3599. doi: 10.15252/embj.201798004. Epub 2017 Dec 7. PMID: 29217591; PMCID: PMC5730888.
- Wu Z, Nicoll M, Ingham RJ. AP-1 family transcription factors: a diverse family of proteins that regulate varied cellular activities in classical hodgkin lymphoma and ALK+ ALCL. *Exp Hematol Oncol*. 2021 Jan 7;10(1):4. doi: 10.1186/s40164-020-00197-9. PMID: 33413671; PMCID: PMC7792353.
- Xiong JW, Leahy A, Lee HH, Stuhlmann H. Vezf1: A Zn finger transcription factor restricted to endothelial cells and their precursors. *Dev Biol*. 1999 Feb 15;206(2):123-41. doi: 10.1006/dbio.1998.9144. PMID: 9986727.
- Ye Zheng, Kami Ahmad, Steven Henikoff. CUT&Tag Data Processing and Analysis Tutorial. 2020 Mar 13. https://yezhengstat.github.io/CUTTag_tutorial/
- Yu W, Ginjala V, Pant V, Chernukhin I, Whitehead J, Docquier F, Farrar D, Tavoosidana G, Mukhopadhyay R, Kanduri C, Oshimura M, Feinberg AP, Lobanenko V, Klenova E, Ohlsson R. Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation. *Nat Genet*. 2004 Oct;36(10):1105-10. doi: 10.1038/ng1426. Epub 2004 Sep 7. PMID: 15361875.
- Zhang Y, Sun Z, Jia J, Du T, Zhang N, Tang Y, Fang Y, Fang D. Overview of Histone Modification. *Adv Exp Med Biol*. 2021;1283:1-16. doi: 10.1007/978-981-15-8104-5_1. PMID: 33155134.

- Zhou Y, Kurukuti S, Saffrey P, Vukovic M, Michie AM, Strogantsev R, West AG, Vetrie D. Chromatin looping defines expression of TAL1, its flanking genes, and regulation in T-ALL. *Blood*. 2013 Dec 19;122(26):4199-209. doi: 10.1182/blood-2013-02-483875. Epub 2013 Nov 7. PMID: 24200685.
- Zou Z, Ocaya PA, Sun H, Kuhnert F, Stuhlmann H. Targeted Vezf1-null mutation impairs vascular structure formation during embryonic stem cell differentiation. *Arterioscler Thromb Vasc Biol*. 2010 Jul;30(7):1378-88. doi: 10.1161/ATVBAHA.109.200428. Epub 2010 Apr 29. PMID: 20431070; PMCID: PMC2903440.