Computational studies of protein interactions and genetic regulation

Joseph Ward

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

Faculty of Biological Science

University of Leeds

February 2013

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

Firstly, I would like to thank David Westhead for his support and guidance over the course of my PhD and for agreeing to be my supervisor. I would like to thank Richard Jackson for his advice and guidance, professionally and personally, during the start of my PhD.

I would like to thank Sean Killen and Aidan Richmond for their help with any and all of my computing problems. My thanks go to the Jackson and Westhead lab members, past and present, including but not exclusive to: Jon Fuller, Sarah Kinnings, Joel Dockray, Carlos Simoes, Nick Burgoyne, Rich Gamblin, James Dalton, Phil Tedder, Tom Forth, Matt Care, Lucy Stead, Al Radford, Salam Assi, Vijay Baskar, Binbin Liu, Mike Bennet and Lee Hazelwood.

I would like to thank the entirety of my family for the help and support they've given me over the course of my PhD. I am also thankful to the Cowley family for accepting me as one of their own. Without the support of both of these families I would never have made it this far. I'd like to thank my friends, in particular the Bryan Wagner-Adair, Tom Benians, Matthew Cowley and Richard Thomas, for their distractions and inspiration in equal measure. A massive thank you goes to Gwen Wagner-Adair for her proof reading skills.

Funding for this studentship was provided by the BBSRC, for which I am grateful.

Lastly, I would like to thank Leanne Cowley, for dragging me out of the lows and standing with me during the highs. Without her, none of this would ever have happened.

This thesis is dedicated to the memory of Steven Charles Cowley.

Abstract

The work in this thesis is split into two parts. The introduction and following two chapters pertain to the investigation of gene regulation using Chip-seq data and linear modelling. The final chapter pertains to the prediction of hot-spot residues in protein-protein interactions.

The rapid escalation in the speed and quality of DNA sequencing has lead to a wealth of data for the location of transcription factor binding and histone modifications across the genomes. Using Chromatin ImmunoPrecipitation followed by sequencing (ChIP-seq) data, we have generated a new binding metric based on the enrichment of the read-counts for each gene.

Eight datasets from mouse macrophage cells (two histone modifications, five transcription factors, DNase I hypersensitivity) were used to model the binding of RNA polymerase II. It was found that a linear model just using the DNase I hypersensitivity and histone modification data was better than any of the models containing the transcription factor data. Investigation of the outlying genes for the model revealed no pattern in their Gene Ontology terms or macrophage-specific genes.

Human embryonic stem cell data (23 transcription factor and 24 histone modification datasets) were used in combination with LASSO regression to model the binding of RNA polymerase II. The resultant models contrasted with the results from the mouse macrophage linear models in that using the histone modifications data in combination with the transcription factor data lead to the best models. A much more complicated picture of the regulation of RNA polymerase II binding was produced using the LASSO models.

Protein-protein interactions are essential for every function within a cell and being able to predict them has large consequences for drug discovery and understanding the vast protein-interaction networks that occur within cells. Predicting protein-protein interactions is difficult due to the large number of possible conformations; predicting hot-spot residues can greatly reduce this. InterBasePro was compared with experimental data and subsequently adaptation was done to assess its usefulness for predicting hot-spot residues. An alternative approach was also made into classifying hot-spot residues based on atomic contacts.

Contents

1 Introduction **1**

1.1 The Genome and Transcription 2

1.1.1 Organisation of DNA 2

1.1.2 Histones 2

1.1.3 Transcription 4

1.1.4 Gene Ontology 11

1.2 Sequencing the Genome 11

1.2.1 Sanger Sequencing 12

1.2.2 Current Generation Sequencing 12

1.2.3 Chromatin Immunoprecipitation 13

1.2.4 DNase-seq 14

1.2.5 Alignment and peak calling 14

1.2.6 Previous Computational Approaches 16

1.2.7 Importance of understanding genetic regulation 19

2 Genetic Regulation in Mouse Macrophage Cells **21**

2.1 Introduction 22

2.1.1 Important Transcription Factors 23

2.1.2 Aims & Objectives 25

2.2 Methods 26

2.2.1 Data sets 26

2.2.2 Enrichment Calculations 26

2.2.3 Linear Models 27

2.2.4 Outlier and Inlier analysis 28

2.2.5 Drop Analysis 31

2.3 Results 33

2.3.1 Linear Models . . . 33
2.3.2 Analysis of the Inliers and Outliers . . . 40
2.3.3 Drop Analysis on linear model . . . 41
2.3.4 Stratification . . . 46
2.4 Discussion . . . 49

3 Genetic Regulation in Human Stem Cells 52
3.1 Introduction . . . 53
3.1.1 Importance of embryonic stem cells . . . 53
3.1.2 Transcriptional control of stem cells . . . 53
3.1.3 Histone modification control of stem cells . . . 54
3.1.4 ENCODE . . . 56
3.1.5 Aims and Objectives . . . 56
3.2 Methods . . . 57
3.2.1 Data sets . . . 57
3.2.2 Enrichment Calculation . . . 64
3.2.3 Linear Models . . . 64
3.3 Results . . . 66
3.3.1 Comparison of Cell Lines . . . 66
3.3.2 LASSO regression . . . 71
3.3.3 Discussion . . . 77

4 Hot-Spot Prediction in Protein-Protein Interfaces 80
4.1 Introduction . . . 81
4.1.1 Properties of the Interface . . . 81
4.1.2 Hot-Spots . . . 84
4.1.3 Experimental Analysis of Protein-Protein Interactions . . . 85
4.1.4 Computational Approaches . . . 85
4.1.5 Binding Site Detection . . . 86
4.1.6 Hot-spot prediction . . . 88
4.1.7 InterbasePro . . . 89
4.1.8 CAPRI . . . 93
4.1.9 Computational Hot-Spot Prediction Methods . . . 94
4.2 Method . . . 96
4.2.1 Adaptation of InterbasePro . . . 96

4.2.2 Alanine Scanning Mutagenesis Dataset 97
4.2.3 Predicting Pockets and Atom Contact Counts 99
4.2.4 Atom Contact Data 100
4.3 Results & Discussion 101
4.3.1 Comparison of Calculations from InterBasePro with Experimental Results from ASEdb 101
4.3.2 Comparison of Calculations from InterBasePro with Experimental Results from the Second Dataset 104
4.3.3 Adapting InterbasePro 105
4.3.4 Atom Contact Data 109
4.4 Conclusions 112

5 Summary **115**

List of Figures

1.1 The first stages of transcription 1: TBP is recruited to the promoter region, changing the conformation of the DNA and allowing TFIID to bind. 2: The general transcription factors and RNA polymerase II (RNAPII) are recruited by TFIID to the promoter complex. 3: TFIIH uncoils the DNA allowing RNA polymerase II to bind fully and access the coding strand of DNA. 4: The majority of the general transcription factors disassociate and RNA polymerase II moves from the promoter region. TFIID remains bound to recruit more general transcription factors and repeat the cycle. 6

2.1 The three regions used for the enrichment calculation. 27
2.2 Binding motifs from JASPAR used in the prediction of binding sites. 32
2.3 The optimal linear model for the logged enrichment of the flanking and gene data. The optimal model predicted RNA polymerase II binding using DNase I hypersensitivity, H3K4Me3 and H3K4Me1 presence. 37
2.4 The residuals of the linear model predicting the log enrichment values for RNA polymerase II enrichment using H3K4Me3, H3K4Me1, DNase I hypersensitivity, CEBPα, p65 and, Pu1. 38
2.5 Residuals for only the macrophage-specific genes for the optimal logged enrichment model using the gene and flanking regions 39
2.6 Comparison between the predicted transcription factor binding sites the genes that were affected the most and least when CEBPα, p65, and Pu1 were removed from the best transcription factor-containing linear model. 43
2.7 Comparison between the predicted transcription factor binding sites for genes that were affected the most and least when pairs of transcription factors were removed from the best transcription factor-containing model. 44

2.8 Comparison between the predicted transcription factor binding sites for genes that were affected the most and the least when all three transcription factors were removed from the best transcription factor-containing model. 45
2.9 Distribution of the log enrichment values for the RNA polymerase II dataset. . . . 47

3.1 Comparison of the correlation between the enrichment scores for three cell lines; H1, H9 and IMR90. 67
3.2 Comparison of the correlation between the enrichment scores for the H1 and H9 cell lines. 68
3.3 Heatmap showing the correlation between the enrichment values for the histone markers for the H1 human embryonic stem cell line. 69
3.4 Heatmap showing the correlation between the enrichment values for the transcription factors for the H1 human embryonic stem cell line. 70
3.5 The linear models for the four H1 embryonic stem cell RNA Polymerase II datasets. 71
3.6 The log linear models for the four H1 embryonic stem cell RNA Polymerase II datasets. 72
3.7 Heatmap showing the LASSO models for the four RNA polymerase II datasets. . 75

4.1 A contrast between interface sizes for protein-protein interactions. The catalase dimer (PDB ID 4CAT) loses 10,570Å^2 of solvent accessible surface area when the two protein chains form an interactions. The superoxide dismutase dimer (PDB ID 1SRD) in comparison loses only 670Å^2 of solvent accessible surface area when the two protein chains interact. 82
4.2 When a hydrophobic molecule is in water, the water molecules form a cage around it. This cage cause a reduction in the entropy of the system. When two molecules move together, the cage that is formed around them is needs fewer water molecules than the two individual cages, this raising the entropy of the system again. 83
4.3 For each atom in a residue, any atoms of the opposing protein chain within 6Å are counted and regarded as having an interaction. A pseudo-alanine scan is performed by counting the number of interactions that are lost if the R-group was removed. In this case 4 of the possible 7 interactions are below the 6Å threshold. 100

4.4 A comparison of the experimental alanine scanning ΔΔG from ASEdb and predicted ΔΔG from InterBasePro, scaled by electrostatics and van der Waals forces using a linear model. 103
4.5 A comparison of the experimental alanine scanning ΔΔG mined from the literature and predicted ΔΔG from InterBasePro, scaled by electrostatics and van der Waals forces using a linear model. 105
4.6 Comparison of the experimental alanine scanning mutagenesis data from ASEdb and predicted energies calculated by InterBasePro. 107
4.7 Comparison of the experimental alanine scanning mutagenesis data from ASEdb and predicted energies calculated by InterbasePro with an added desolvation term based on the octanol-water transfer energy of amino-acids 108
4.8 Comparison of the atom contact count lost when a residue is mutated to alanine and the experimental energies from ASEdb. 110

List of Tables

2.1 The datasets used for the epigenetic analysis of mouse macrophage cells 26

2.2 A list of genes known to be important to the development or function of mouse macrophage cells taken from Ravasi *et al.*. 30

2.3 Analysis of the Linear models containing different combinations of the predictors for the flanking regions and gene region logged enrichments. 34

2.4 Analysis of the Linear models containing different combinations of the predictors for the Gene region logged enrichments. 35

2.5 Analysis of the Linear models containing different combinations of the predictors for the flanking regions logged enrichment. 36

2.6 The over-represented GO terms from the genes with the 5% largest and 5% smallest for the linear model predicting the logged enrichment values of RNA polymerase II using H3K4Me3, H3K4Me1, DNase I hypersensitivity, p65, Pu1 and CEBPα logged enrichments as predictors. 41

2.7 Models for the stratified data using various groupings of genes. 46

3.1 The datasets used for the generation of models for human H1 embryonic stem cell line from the GEO database. 61

3.2 The datasets used for the generation of models for human H9 embryonic stem cell line from the GEO database. 62

3.3 The datasets used for the generation of models for human IMR90 embryonic stem cell line from the GEO database. 63

3.4 Summary of the LASSO models for the four RNA polymerase II datasets for the enrichment values and logged enrichment values. 73

3.5 Summary of the LASSO models for sub-sets of the four RNA polymerase II datasets for the enrichment values and logged enrichment values. 73

4.1 Octanol-water transfer energies used as solvation factors for InterbasePro 97

4.2 A summary of the dataset used from the Alanine Scanning Experiment Database 98
4.3 The alternative dataset for the comparison of InterbasePro predicted data to alanine scanning mutagenesis data. 99
4.4 Correlation coefficients for the comparison of InterBasePro predicted changes in binding energy with the alanine scanning mutagenesis data from ASEdb. . . . 102
4.5 Correlation co-efficients for the comparison of InterBasePro predicted changes in binding energy with the experimental alanine scanning gathered from the literature. 104
4.6 Sensitivity and specificity for the linear model scaled predicted values from InterBasePro compared with the ASEdb and alternative experimental data sets. . . 105
4.7 Correlation coefficients for the comparison of the predicted energies from InterbasePro and the experimental energies of ASEdb. 106
4.8 Correlation coefficients for the atom count data. Total is the correlation coefficient when regarding all the atoms in each residue. 109

Abbreviations and Acronyms

H2A	Histone 2A
H2B	Histone 2B
H3	Histone 3
H4	Histone 4
H1	Histone 1
bp	Base pairs
kbp	Thousand base pairs
H3K4Me1	Mono-methylation of lysine 4 of histone 3
TFB	Transcription factor binding protein
TFII	Transcription factor II
DPE	Downstream promoter element
CEBP	Ccaat-enhancer binding protein
GO	Gene ontology
DNA	Deoxyribose nucleic acid
RNA	Ribonucleic acid
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
PPI	Protein-protein interaction
ASEdb	Alanine scanning experiment database
SASA	Solvent accessible surface area
PDB	Protein data bank
UCSF	University College San Fransisco
NMR	Nuclear magnetic resonance
CAPRI	Critical Assessment of Prediction of Interactions
BID	Binding interface database
INF1	Interferon-1
TNF	Tissue necrosis factor
IL-10	Interleukin 10
BCL6	B-cell lymphoma protein 6
LPS	Lipopolysaccharide
AML	Acute myeloid leukemia
CREB	cAMP-responsive element-binding protein
NCBI	National Center for Biotechnology Information
GEO	Gene Expression Omnibus
BIC	Bayesian information criterion
ChIP	Chromatin immunoprecipitation
BP	Biological process
MF	Molecular function
CC	Cellular component
ENCODE	Encyclopedia of DNA Elements
LARS	Least Angle Regression

Chapter 1

Introduction

This thesis describes two different projects. This introduction and first two chapters are centered around using high-throughput DNA sequencing data to model genetic regulation with the aim of investigating a mechanistic hypothesis for RNA polymerase II binding. The final chapter, with its own self-contained introduction, centers around the prediction of hot-spot residues in protein-protein interfaces.

1.1 The Genome and Transcription

1.1.1 Organisation of DNA

Human DNA from a single cell is comprised of about 3 billion bases. If stretched out, this would be over 2 meters long. Without being highly organised this would, quite simply, not fit into the nucleus of every cell. The double-helical strand of DNA is first organised by being wrapped around a histone protein complex. This complex is made up of 8 proteins, two H2A histone proteins, two H2B histone proteins, two H3 histone proteins and two H4 histone proteins. Each histone protein complex has about 200 base pairs of DNA wrapped around it. The coiling of DNA around the histones reduces about 680Å of DNA into a 55Å by 110Å cylinder.[1] This is the first stage of the packing of DNA and is known as a nucleosome. The nucleosomes are linked together by the H1 histone protein. This pulls the complexes closer together where they form an incredibly dense coil. In metaphase, the stage of mitosis where the cell is splitting into two and the majority of the DNA is inactive and tightly packed, with a total chromosomal length of approximately 200μm. This is a packing ratio of 10^4. As incredible as this organisation is, the tight packing of the nucelosomes and the coiling of the DNA around the histone complexes means that the DNA cannot be easily bound by proteins or other mechanisms needed for it to be transcribed to RNA.[1]

1.1.2 Histones

Histone proteins are the core of the organisation of DNA. Histones all have a high ratio of positively charged amino-acid residues; approximately one in four residues are arginine or lysine. This high ratio of positively charged residues allows the histones to bind well to the negatively charged phosphate backbone of the DNA strands. The essential function of these proteins means that they are incredibly highly conserved. The H3 and H4 histones are the slowest evolving of all eukaryotic genes.[2] Interestingly, the H3 histones present within centromeres

have a highly conserved core across many species but also have a massively divergent N-terminal tail.[3]

Histone proteins are highly toxic to cells and as such their expression is tightly controlled. Problems arise during mitosis, however, as DNA is duplicated at a massive rate, during which a large amount of histone proteins are rapidly needed to pack and regulate the newly-produced DNA. The regulation of histone protein transcription is still poorly understood. Interestingly, it is thought that there are potentially two master regulators of histone genes, E2f1 and E2f4.[4] It was also found that there are specific regulators for the core (CTCF) and linker (Zfx) histones as well as cell-type specific regulators.[4]

Histone Modifications

Histone modifications are essential in several ways for controlling the transcription of DNA. They control how tightly the DNA helix is associated with the histone complex, and they can have a direct effect on the binding ability of transcription factors and the transcription initiation complex.

Acetylation of lysine in histones is generally associated with the activation of transcription. Acetylation, by histone acetylation transferases, of the histone leads to a reduction in the positive charge of the histone. This means that the negatively charged DNA is less-tightly associated with the histone complex, making it more accessible for transcription.[5] De-acetylation, by histone deacetylases, can also be employed by signalling pathways as a repressor of transcription as they increase the binding of the DNA to the histone protein complex.[6]

Methylation of lysine and arginine residues in the histone proteins has a wide variety of functions from large-scale re-modelling to activation and repression. H3K4 methylation is correlated with transcription of a gene. H3K4 mono-methylation normally peaks near the end of the transcribed region, while H3K4 di-methylation peaks near the center of the transcribed region and H3K4 tri-methylation forms a sharp peak near the transcription start site.[7] While H3K4 is correlated with transcription, the exact mechanisms by which H3K4 methylation encourages transcription are not fully known.[8;9] Unlike histone acetyltransferases and histone deactylases, lysine methyltransferases are normally specific to a single lysine residue on a single histone.[10] H3K36 and H3K79 methylation have both also been associated with transcriptional activation. Methylation of H3K9, H3K27 and H4K20 are associated with repression of transcription. The exact mechanisms of repression by histone methylation are not clear for all markers, but H3K36 methylation is known to recruit histone deactylases. The deactylation of the histone complex results in a lower propensity for the associated DNA to be transcribed.[11;12] While methylation

of some lysine residues is used to repress transcription of genes, the demethylation is also used as a process of activating transcription. LSD1 demethylation of H3K9 has been shown to lead to transcription of the associated gene.[13]

The overall picture of histone modifications is under intense investigation. The simple picture outlined here is by no means comprehensive. The intricate interplay between the vast array of histone modifications cannot identified using the current detection methods. High-throughput techniques can only identify the presence of a single modification at a time and do not identify the vast array of markers that are all potentially present on a single histone complex.

1.1.3 Transcription

Transcription, the process of decoding the sequence of DNA bases in an RNA molecule, is the first step in the production of proteins. Transcription is a highly regulated process which can be split into five separate steps: pre-initiation, initiation, promoter clearance, elongation and termination. The key protein complex responsible for transcription is RNA polymerase II. The RNA polymerase II complex is responsible for the recruitment of appropriate RNA molecules to the DNA sequence. There are three types of RNA polymerase enzymes: RNA polymerase I, II and III. RNA polymerase I is responsible for the production of 18S, 5.8S and 28S ribosomal RNA molecules. RNA polymerase III is responsible for the production of 5S ribosomal RNA and of recruiting the correct amino acid during translation. RNA polymerase II is responsible for transcribing the mRNA molecules which for encode proteins and for other RNA molecules, such as snRNA.

Pre-initiation and initiation

Pre-initiation is the first stage of transcription and involves the binding of general transcription factors to the promoter of the gene. In eukaryotes, one of the most understood core promoter elements for genes is the TATA-box that resides 10 to 25 base pairs upstream from the transcription start site. The pre-initiation stage involves the binding of the transcription factor II D (TFIID) protein complex, which includes the TATA-binding protein (TBP). When TBP is bound, the DNA bends to an angle of 80° towards the major groove, optimising the orientation of the DNA for the binding of TFIID.[14] The pre-initiation stage is highly controlled and continues with the binding of a protein complex centered around TFIID.[15] After TFIID has bound to the TATA-box, TFIIA binds to the upstream edge of TFIID and TFIIB binds to the downstream edge of

TFIID (see figure 1.1). The binding of TFIIB recruits the RNA polymerase II complex to the DNA and the location of this determines the exact start site of transcription, as well as the direction of transcription.[16–19] TFIIF is recruited at this stage and causes a conformational change in the initiation complex. This conformational change stabilises the complex and causes the DNA to wrap around the RNA polymerase II complex. TFIIE then binds to the downstream edge of RNA polymerase II and enhances the association of RNA polymerase II to TFIIB.[20–22] The binding of TFIIE promotes RNA polymerase II to begin separating the DNA strand and recruits TFIIH to the pre-initiation complex. TFIIH is the final part of the pre-initiation complex and has two functions; and ATP-helicase and kinase activity. TFIIH is essential for opening the DNA and initiating transcription.[23–25]

With the binding of TFIIH, with its helicase and kinase functions, transcription can be initiated. The helicase function of TFIIH unwinds the DNA so that a single strand is accessible to the RNA polymerase II. The RNA polymerase II complex then moves off from the promoter region, leaving TFIIA and TFIID on the promoter.

The TATA-box is not the only core promoter element that can be present. The initiator promoter element (INR, facilitates the binding of TFIID) can be present on or next to the transcription start site, the TFIIB recognition element (BRE, facilitates the binding of TFIIB) may be present adjacent to the TATA-box and the downstream promoter element (DPE) may be present 30 bases downstream from the transcription start site.

Not all genes have a TATA-box for the pre-initiation complex to bind to. Genome wide studies have shown how little is known about these promoter motifs. Only 10% of human promoters were found to have a canonical TATA-box and 25% to have a TATA-like motif.[26;27] Only 46% of human promoters were found to have a INR-motif with 30% of the total number of promoters having INR-motifs but lacking a TATA-like motif.[27] TATA-binding protein, when not associated with TFIID, is highly specific for the TATAA motif. When TATA-binding protein is bound to TFIID, however, TFIID loses this specificity.[28–30] TFIID binding is still required for the initiation of transcription, regardless of the presence of a TATA motif.[20] There is still much work to be done in identifying promoter motifs and the vast amount of data produced by the current generation of sequencing methods is helping to identify new motifs.[27;31]

Promoter clearance, elongation and termination

After the RNA polymerase and transcription factor binding complex has associated to the DNA, RNA polymerase II has to get clear of the promoter before it can begin transcribing the DNA to RNA. Initially RNA polymerase II will go through a phase of “abortive transcription”; a process

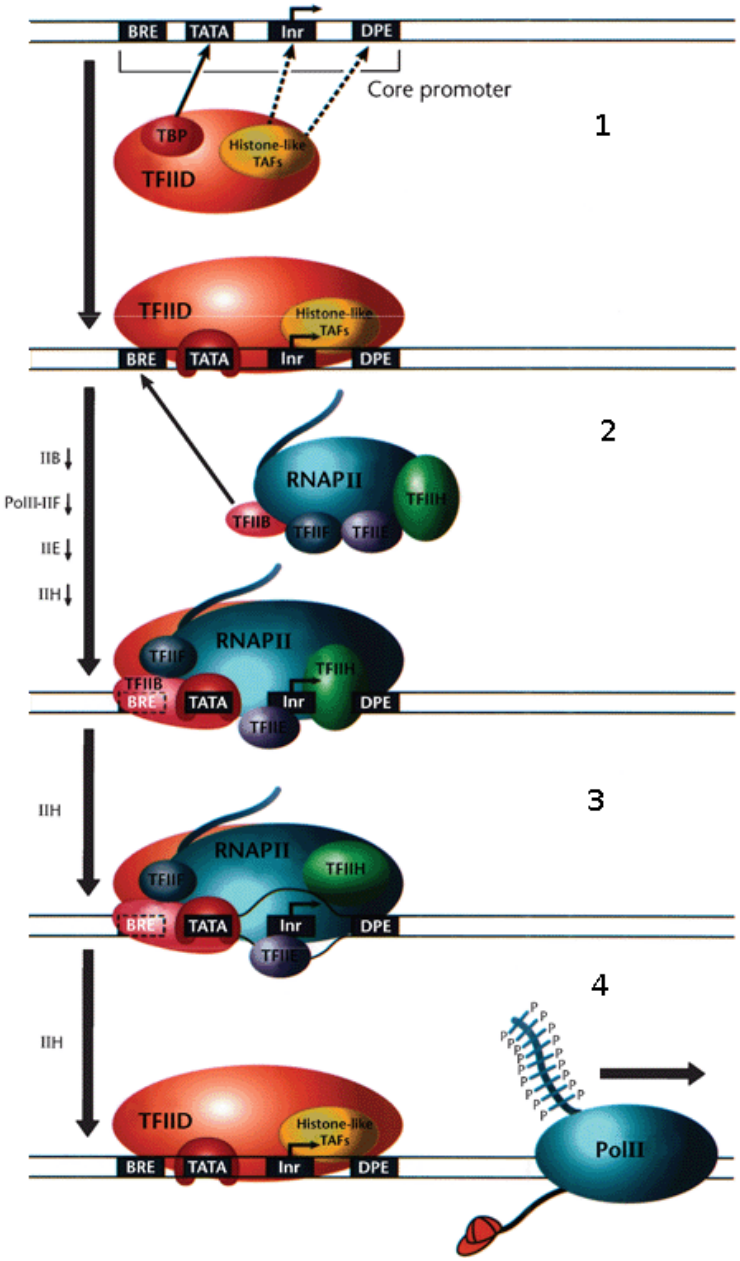


Figure 1.1: The first stages of transcription 1: TBP is recruited to the promoter region, changing the conformation of the DNA and allowing TFIID to bind. 2: The general transcription factors and RNA polymerase II (RNAPII) are recruited by TFIID to the promoter complex. 3: TFIIH uncoils the DNA allowing RNA polymerase II to bind fully and access the coding strand of DNA. 4: The majority of the general transcription factors disassociate and RNA polymerase II moves from the promoter region. TFIID remains bound to recruit more general transcription factors and repeat the cycle. TAF: TBP-Associated Factor. Figure modified from Nicolas *et al*[14]

where the first 2-15 bases are transcribed but then the RNA is released and the transcription complex dissociates from the DNA or returns to the promoter.[32] This continues until TFIIH phosphorylates serine-5 of RNA polymerase II, allowing it to progress past the promoter region.[33] Once elongation is established, RNA polymerase II unwinds the DNA by two turns of the helix. Methylation of lysine 36 of the H3 histone protein is essential for the progression of RNA polymerase II. As the RNA is transcribed, a 12bp of RNA-DNA hybrid with just under two turns of the DNA helix. There is a sharp kink in the RNA at the end of this DNA-RNA hybrid helix to ensure the RNA does not interfere with the two DNA strands as they re-associate. Once the RNA strand reaches a length of 25 nucleotides, the transcription factors that were originally needed for the recruitment of RNA polymerase II dissociate leaving the RNA polymerase II to carry on transcribing the DNA. At this point the RNA molecule being synthesised is capped. The cap is a tri-phosphorylated guanine nucleotide that defies the canonical 5'-3' phosphate link of normal nucleotides in that it forms a 5'-5' tri-phosphate link. Capping is essential for the formation of mature RNA; without the cap the RNA is susceptible to 5' endonucleases and degraded.

The process of transcription is a re-iterative process. The transcription complex moves along the DNA one base at a time. The appropriate base-matching nucleotide is recruited to the active site of the complex and a condensation reaction, catalysed by two magnesium cations, results in the extension of the RNA molecule by one base and the release of pyrophosphate. Translocation moves the transcription complex forward a single base to re-expose the active site and the process continues. The 12bp DNA-RNA hybrid is always maintained at the same length; when a base is added to the 3' end of the RNA molecule a base dissociates from the DNA at the 5' end. The mechanism that limits movement of the transcription factor complex by a single base each step is still unknown and currently has not been successfully modelled using molecular dynamics.[34] The transcription complex is able to move backwards a single base along the DNA. This generally occurs when an incorrect base is incorporated into the RNA so that it can be removed.

Termination of transcription is as controlled as the previous stages and is a dynamic process. Poly(A)-dependent termination requires a 5'-AAUAAA-3' RNA sequence followed by a G/U rich sequence being transcribed. Upon the AAUAAA sequence leaving the transcription complex, CPSF, a protein bound to RNA polymerase II, binds to the RNA and pauses transcription. G/U rich sequence that follows is then bound by CtsF and promotes the cleavage of the RNA by CPSF between these two signals. The resulting upstream RNA molecule is polyadenylated and completes its maturation while the downstream product is degraded by

XRN2. It is thought that when XRN2 encounters RNA polymerase II when it is degrading the RNA that it promotes the release of RNA polymerase II from the DNA and helps the final stage of termination.[35] Interestingly, the genes for histone proteins do not use a Poly(A)-dependent mechanism for termination but rely on a stem-loop mechanism instead.[36] Cleavage of the stem-loop-dependent termination is similar to that of Poly(A)-dependent termination in that is occurs between two elements, the stem-loop and a purine-rich histone-downstream element, and it is cleaved by CPSF[36]. The stem-loop is a pair of sections of RNA with a short linker between them. The two sections of DNA bind to each other to form the stem, leaving the linking section as a loop. The current knowledge of transcription termination is a rapidly developing field, so much so that four reviews have been published in the last three years; each addressing the new knowledge that has been accumulated.[37–40]

Enhancers and Repressors

Transcription as described so far is the general case and results in a low basal level of transcription. The genes need to be enhanced to increase transcription beyond this background level. To reduce or stop transcription, genes need to be repressed. Specific transcription factors, as opposed to the general transcription factors so far described, are responsible for each of these roles and this is how cells exert control over the specific genes that are transcribed to determine cell fate, function and respond to extra- and intra-cellular signals.

Enhancer regions are sections of DNA, normally about 100 bases long, which are bound by transcription factors that promote transcription. They contain several short 6-12 base transcription factor binding motifs. These motifs have some degree of conservation that the transcription factors can recognise. There are many different mechanisms by which enhancers and their associated transcription factors function. Co-operative binding occurs where transcription factors associate with each other to enhance transcription. An example of this is the recruitment of p300 to the enhancer of INFβ . p300 associates with NFκB, IRF1 and c-Jun.[41] Individually they do have an effect on transcription but together the effect is amplified. This often occurs where the transcription factor interacts with the transcription initiation complex. Co-operative enhancers can occur at a single site; two transcription factors can, counter intuitively, promote transcription despite competitively binding to the same binding site. It is thought that this occurs as the transition between the two transcription factors leaves less opportunity for nucleosomes to re-associate with the open DNA.[42]

The binding of a transcription factor can cause local bending in the DNA. The change in the conformation of the DNA can open up enhancers to transcription factor binding or open

promoters to the initiation of the transcription initiation complex. The bending of DNA can also bring transcription factors into the proximity of the transcription initiation complex, enhancing its propensity to undergo promoter clearance.[43;44]

Transcription factors are compete for binding sites on the DNA with nucleosomes. Binding by transcription factors such as Pu1[45] and CEBPβ[46] to their enhancers leads to remodelling of chromatin and the opening of the local DNA to the binding of other transcription factors. This is the first stage in the differentiation of the cells to several different possible cell-types. The effect that transcription factor binding has on nucleosomes is complicated. The binding of transcription factors is dependent on the chromatin state but the chromatin state is also be dependent on the transcription factors that are bound.

The organisation of the motifs also has an important effect on the function of the enhancers and their associated transcription factors. The order of the enhancers on the genome determines which transcription factors interact; changing the order of the motifs moves the interface between two proteins so that they are no longer able to interact. Changing the direction of the motif can have the same affect as it changes the orientation in which the transcription factor binds.[47;48] If the distance between two motifs changes, the location of the two transcriptions factors around the DNA helix will be changes, also disrupting the interaction of the two proteins. No general pattern has been found between the order, orientation or separation of transcription factor binding motifs.[49]

Repression of transcription shares many similarities with the enhancement of transcription. Many of the mechanisms, such as co-activation, are the same for repression but instead of the bound transcription factor promoting the binding of the pre-initiation complex, or the progress from initiation to elongation, it reduces or completely inhibits them.[50] Repression is used as a fine-grained control over the levels of transcription. Oct4 is an essential transcription factors that not only maintains pluripotency in embryonic stem cells but also determines cell fate. Different levels of Oct4 determine whether a cell continues as a stem cell or begins differentiation, and what it will differentiate to.[51;52] Oct4 expression is complicated and not fully understood. Cdx2 has been shown to repress the transcription of Oct4, but it requires the presence of other factors such as Brg1.[53] Repression is as tightly controlled as enhancement of transcription and often requires multiple-levels of transcription factor binding. Repression also allows cells to maintain proteins in a poised state. This poised state means that the transcription initiation complex is bound but never progressed to the elongation phase. This allows a rapid response to signalling and allows the cell to very quickly start transcription of the poised gene.[50;54;55]

This is a simplified overview of enhancement and repression of transcription. It has been

known for a long time that the function of a transcription factor is not always clear and that they can both repress and enhance the transcription of genes depending on the surrounding environment.[56–58] Establishing the function of transcriptional enhancers and repressors is difficult. Transcription enhancer regions are often located remote to the promoter and transcription start site. Since DNA functions in a three-dimensional environment these can be a large distance between the enhancer and the gene it enhances. The multiple layers of interactions also makes it harder to understand a single transcription factors effect. The vast networks of interactions lead to different results dependent on which other proteins are present. It has only been recently, with the development of new sequencing technologies, that is has been possible to get a whole-genome over-view of the location of individual transcription factors on the genome and the investigation of these complicated processes.

Predicting Transcription Factor Binding Sites

Much work has been done on the prediction transcription factor binding motifs; the short set of nucleotides that each transcription factor binds to. Transcription factors are specific in the motifs which they bind to; incorrect binding could have disastrous consequences for the cell. This is advantageous for the prediction of the transcription factor binding motifs as it means they are often highly conserved. There are publicly available databases, such as JASPAR,[59] and TRANSFAC[60] that contain experimentally determined binding motifs for transcription factors. The motifs generally consist of a position weight matrix with the probability of the occurrence of each base at each location in the motif. These matrices can then be combined with methods, such as MOODS, to identify potential transcription factor binding sites.[61] MOODS searches through the DNA sequence for the most likely sub-string, determined from the position weight matrix, to identify potential binding sites. After these core sections have been identified, they are assigned a p-value based on the whole of the position weight matrix.[61]

Despite the work that has been done in predicting transcription factor binding sites using motifs it is still a difficult and largely unsolved problem. The methods produced, though capable of providing predictions of whether a motif is present or not, are not always reliable. Due to the small size of the motifs ($\sim$6bp) and each base being only 1 of 4 nucleotides there is a high probability that the motifs can occur by chance. A 6-base motif would be expected to be seen, by chance, once in 4096 bases. With a genome as large as that of humans, approximately 3 billion bases, naively assuming a random distribution of bases, there would be, by chance, over 700,000 occurrences of each 6-base motif. The short length of the motifs mean that there they do occur by chance in the human genome. This is problematic for predicting transcription

factor binding sites just using position-weight matrices; the methods currently available lack the context that occur in the cell to determine whether, at that specific location in the genome, the sequence is a valid binding site. There are numerous other problems, beyond the statistics, that make predicting binding sites problematic. The poorly understood biology behind the binding of many transcription factors, variability in the affinity of binding and the variability in the sequence a protein will bind to all add to the complexity of the problem. Even more complexity is added when the variation of each base in the binding motif is not assumed to be independent of every other base, as is the assumption when using position weight matrices. It has been show that often the variation in the bases is dependent on the variation in the other bases of the motif.[62] Position weight matrices are still commonly used however, due to their relative simplicity.

1.1.4 Gene Ontology

The Gene Ontology (GO) is a collaborative project which aims to provide a consistent description of gene function across a number of different organisms. It contains 3 sub-ontologies that describe the associations of the gene product: Cellular Component, Molecular Function, and Biological Process. The Cellular Component ontology describes the locations of the gene product in the cell or extracellular matrix where it is often found (e.g. the cytoplasm or inner membrane). The Biological Process ontology describes the molecular events or process with which the gene product is associated. The Molecular Function ontology describes the molecular activities of the gene product. The terms are designed to be simple, concise and descriptive in order to minimise the amount of time and effort needed to identify the function, location and actions of a gene product.

1.2 Sequencing the Genome

The increase in the rate at which it is possible to sequence DNA is increasing every year . What originally took the human genome project years to complete is now possible in days, potentially hours. The rapid increase in the speed of sequencing also brings with it a massive wealth of data. Studies using vast ranges of binding data from transcription factors and histone modifications are now possible. As the sequencing technologies become quicker, cheaper, and easier, the function and interplay between all of the regulatory units of transcription and cellular function are going to become clearer.

1.2.1 Sanger Sequencing

Developed in 1975 by Frederick Sanger,[63] Chain-termination, or Sanger Sequencing, was one of the main methods for sequencing DNA, prior to the newest generation of sequencing methods. Sanger sequencing uses modified nucleotides, dideoxynucleotides, to terminate the extension of DNA replication. Single-stranded DNA is separated into four separate reactions. Each reaction has a full range of normal deoxynuceleotide, a DNA polymerase, and a single type of dideoxynucleotide. The DNA is allowed to extend; where the dideoxynucleotide is used the extension is terminated. This results in chains of different lengths. All four reactions are then heat denatured and run through gel electrophoresis. These different lengths from the reactions containing each dideoxynucleotide can then be read off the gel, giving the sequence of the DNA. This method can sequence DNA strands of up to 1000 bases but it is expensive and time-consuming.

1.2.2 Current Generation Sequencing

The current generation of DNA sequencing has a massive advantage over Sanger sequencing as it is massively parallel. Sanger sequencing only samples a single section of DNA at a time. The new sequencing methods can sequence millions of reads for each run of the machine. Being able to sequence millions of reads during a single run of the machines means that it is possible to sequence across the whole of a genome, or a smaller sequence of DNA a vast number of times, where previously it would have been only possible to sequence a single, relatively short, section of DNA. Three of the main sequencing methods, developed by three different companies, are 454 Sequencing, Illumina Sequencing and SOLiD Sequencing.

454 Sequencing, like Sanger sequencing, relies on “sequencing by synthesis”, whereby the complementary strand to the target must be made, step by step, and the nucleotide is then detected to determine the sequence. This is done using parallelised pyrosequencing. DNA is immobilised into beads in wells on a plate to act as the template. Nucleosides, specifically deoxynucleoside triphosphate, are added to the wells and flushed off, one at a time. If the current nucleotide is incorporated into the DNA strand, the release of pyrophosphate causes fluorescence which can then be detected; the order of the bases in the template sequence can then be deduced. 454 Sequencing produces reads of about 300-500 bases.[64]

Illumina Sequencing also relies on the synthesis of the complementary strand of DNA to the target strand, but all of the nucleotides needed for the synthesis are included in a single stage. The DNA molecules that have been enriched by ChIP are attached to a slide. The at-

tached DNA is then clonally amplified; multiple copies are created in a small region to make the imaging stage of the process easier. The nucleotides used in this process are reversibly bound to a dye-terminator which fluoresces when bound. Each base fluoresces a different color and inhibits further extension of the DNA chain until the dye-terminator is removed. Illumina Sequencing involves repeated steps to extend the complementary DNA by a base, flushing away of the unused nucleotides, recording the fluorescence, and removing of the dye-terminator from the base. Illumina Sequencing is very fast and produces read lengths of up to 200 bases. The recently released Illumina machines can produce reads up to 500 bases in length, though the accuracy is severely reduced at the far end of the reads.[65]

The third method, SOLiD (Sequencing by Oligonucleotide Ligation and Detection) sequencing, is based on the principle of "sequencing by ligation". Instead of adding a single base and recording which base is added, all possible combinations of 8-base sequences are present and uniquely marked in the reaction mix. These oligonucleotides have a cleavage site between the 5^{th} and 6^{th} base. SOLiD sequencing comprises of 5 rounds, each round having 5-7 cycles. Firstly, a universal primer is bound to the DNA, then the reaction mixture containing the 8-base oligonucleotides is added and the oligonucleotides allowed to bind to the DNA. The reaction mix is then washed off and the fluorescence of the bound oligonucleotides is measured. The oligonucleotide that has bound is then cleaved, removing bases 6, 7 and 8, as well as the fluorescent signal. The process then continues with the addition of the reaction mix. This process is repeated several times, extending the chain and recording the oligonucleotide sequences that have bound. For the next round, a universal primer binds so that it ends one base before the previous primer ended. This means that each time an oligonucleotide binds, the cleaved bases are tested and, due to the shifting of the primer, each base gets tested twice. The read lengths produced by this method are about 200 bases in length.[66]

1.2.3 Chromatin Immunoprecipitation

Chromatin Immunoprecipitation, or ChIP, is a process for selectively enriching short segments of DNA which are bound by target proteins or epigenetic markers, such as a modified histone protein. First, all the proteins bound to the protein are covalently bound to the DNA, normally using formaldehyde. The DNA is then sonicated to break it into small segments. The segments of covalently bonded DNA are then selected for by the addition of beads with the appropriate antibody for the protein bound to them. The beads are then isolated and the final step is completed, breaking the covalent protein-DNA bonds and leaving the isolated DNA[67].

Chromatin Immunoprecipitation and the new sequencing methods are now often combined to form a powerful method for identifying the location of transcription factors or histone modifications across the genome. As ChIP enriches DNA sequences that a target factor bound to, subsequence sequencing of these DNA sequences allows alignment of the resulting sequence to a reference genome. Aligning the sequences to the genomes allows the binding sites of transcription factors or presence of histone modifications to be studied on a genome-wide scale.

1.2.4 DNase-seq

DNase-seq has been used to identify open chromatin across the genome. DNA is cleaved using DNase I. Biotinylated linker I is added to the cut-ends of the DNA. The DNA is then selected and amplified using PCR, followed by high-throughput sequencing normally using the Illumina methodology.[68] DNase-seq is a genome-wide method which allows scientists to analyse how accessible the DNA is to transcription factors and the transcription factor binding machinery; if the DNA is not accessible, then it can not be transcribed.

1.2.5 Alignment and peak calling

Alignment

Next generation sequencing produces many short reads, spread across the genome where target proteins have been bound or histone markers found. The next stage is locating where these reads actually belong on the genome. Where a reference genome sequence is available, it is possible to align the DNA sequences from the sequencing to the same sequence from the reference genome. Two of the most used programs, BWA[69] and BowTie[70] use the Burrows-Wheeler Transform for the alignment. Ideally, a read will only align to a single location along the genome. Issues arise where a read will map to multiple positions in the genome. Where the read can align to multiple places on the reference genome they are given mapping scores. Mapping scores are used in later stages of analysis where reads with a mapping score below a set threshold are discarded.

Peak Calling

Peak calling is a common method for identifying transcription factors and for histone modification sites from ChIP-seq data. The general aim of peak calling is to identify where reads are

significantly enriched compared to the background, and hence to identify where the target transcription factor or histone modification is bound. Generally, to normalise the ChIP-seq data, the data is either compared to a control run of data generated using non-specific antibodies (or the input DNA which has not undergone ChIP) or to multiple other Chip-seq datasets. A simple method used to normalise the data is to scale the total number of reads for each dataset up to the same number as the largest dataset. This is done by calculating a scaling factor (the largest number of reads divided by the number of reads for the current dataset). This is a broad, linear scaling factor that will allow a fair comparison of the number of reads at any point in the dataset. The simplicity of this method poses some problems. The scaling amplifies the valid peaks, but also artificially inflates any erroneous peaks, as well as any background noise. It assumes that there is a uniform distribution of background reads across the dataset, as well as a similar distribution of reads across both of the datasets being normalised. A problem with peak calling algorithms is that they are often only optimised for the detection of sharper transcription factor binding peaks, the broader peaks of histone modifications or RNA polymerase II binding. As transcription factors bind at specific promoter/binding regions that are very small and precise, the manner in which the reads line up over this area cause tall, thin peaks. Histone modifications and RNA polymerase II and act over a much larger length of DNA which causes broader, lower peaks, compared to the peaks cause by transcription factors.

MACS[71] is a highly-used open-source peak finding tool. The first step MACS takes in the process of calling peaks is to estimate the split between the peaks for each binding site. Since DNA is double-stranded, and each strand has an equal chance of being selected and sequenced, a bi-modal peak occurs where the transcription factor of interest binds. MACS calculates the shift-size between the two peaks using a sliding window twice the size of the sonication-fragment; it looks across the genome for these bimodal peaks. MACS then models the reads across the genome using a Poisson distribution. All of the reads are moved by half the calculated shift-size in the 3' direction of the tag to align the bimodal peaks, then slides a windows twice the shift-size along the genome to detect these merged peaks. A dynamic expected number of reads, λ_{local}, is used to calculate the probability of the number of reads in the peak due to chance and any peak with a p-value below 10^{-5} is highlighted as a binding site. λ_{local} is calculated as in equation 1.1.

$$\lambda_{local} = max(\lambda_{BG}, \lambda_{5k}, \lambda_{10k}) \tag{1.1}$$

where λ_{5k} and λ_{10k} are the number of events per interval of a 5 kb and 10kb window centered on the peak, respectively. λ_{BG} is the number of events per interval for the whole genome or the control dataset if available. This dynamic calculation of λ_{local} allows for a robust calculation of the p-value for each peak which accounts for local variation in the background noise.[71]

Another peak calling algorithm, PeakSeq[72], uses three stages in the identification of peaks. Firstly, a signal density map is produced from the uniquely aligned reads on the ChIP-seq data. This involves extending the reads in the 3' direction to the length of the DNA fragments in the sequencing library (normally about 200bp). Once this has been done the signal density map can be generated as the number of reads that overlap each base. In the next step, PeakSeq identifies all of the regions in the sequence density map that are significantly higher when compared to a simulated null background model, in 1Mb segments across the genome. The peaks that are identified as potentially significant are then compared to the control dataset. 10kb windows across the control and sample data are used for linear regression, with the counts for the control data then being scaled by the slope of the regression line to normalise the two datasets. Any potentially-significant peaks are then compared to the normalised control data and any peaks which are significantly enriched compared to the control are highlighted as binding sites. p-values for the peaks are calculated using a binomial distribution.[72]

PeakRanger[73] builds on PeakSeq with the aim of being able to detect the sharper peaks of transcription factor binding and the broader peaks of histone modifications or RNA polymerase II binding. PeakRanger uses the same methods as PeakSeq to identify potential peaks then uses a "summit-valley-alternator" algorithm to identify the highest point in each peak above a certain threshold, calculated by multiplying the height of the current highest point by a user-defined constant. This is effectively looking for peaks followed by troughs, requiring the height of the reads to fall below the threshold before another peak can be called. This method can look for both steep peaks with small bases, and lower, broad peaks, and can also distinguish between peaks that are close together.[73]

1.2.6 Previous Computational Approaches

With the development of high-throughput sequencing data comes a wealth of new data. There have been very few computational studies looking at using either transcription-factor data or histone-modification data to predict expression.

Studies have been done into using transcription factors to predict gene expression. Ouyang,

Zhou and Wong took the approach of using principle component analysis to construct models using 12 transcription factors from mouse embryonic stem cells. Ouyang *et al.* developed a metric that uses peak intensity and the proximity of the ChIP-seq peak to each gene. The influence a transcription factor has on a gene is the weighted sum of the intensities of all of the transcription factors peaks, termed the transcription factor association score. The weight is based on the proximity of the ChIP-Seq peak to the gene and tends to 0 for the peaks further than about 10kb. The decay in the weight for a peak decreases exponentially the further from the gene it is located, meaning that the largest contribution to the association score is from the ChIP-seq peaks closest to the gene. This is calculated for each transcription factor with each gene. They compared using a binary classifier, stating simply whether the transcription factors is bound to the gene or not, and continuous version of their metric and found that using a continuous metric resulted in far better models than using a binary metric.[74] Principle Component Analysis is a regression method that converts a set of data into uncorrelated subsets of the data. Each of these subsets is comprised of a weighted sum of the original datasets that explains a part of the total variation of the observed data. Principle component analysis was used as it allows different combinations of transcription factors to have positive and negative effects on a genes expression depending on what other transcription factors are bound, with each of the different combinations being contained in a separate principle component. While this is good in principle, the first principle component found contained only one transcription factor and accounted for almost half of the variance in the RNA-seq data used as a measure of transcription. Ouyang *et al* did, however, find that their first three principle components accounted for almost 70% of the variance in the RNA-seq data and was able to identify differentially expressed genes in the embryonic stem cells.[74] The transcription factor association score Ouyang *et al* developed starts approaching the problem of distal transcription binding sites; sites that are not necessarily close to the gene but has influence on its transcription. It does not, however, fully encompass the problem as there is still no accounting for long-range distal effects or 3-dimensional folds in the DNA. The approach uses a simplistic idea that the further from the gene a binding site is, the less likely it is to influence the genes expression. While this is a safe general assumption it is quite possible that the cases where this assumption is incorrect, such as enhancer regions that have a strong influence on gene expression, will have a significant effect on the model.

Cheng and Gerstein used support vector regression to investigate the relationship between transcription factor binding, histone-modification presence and expression.[75] Using ChIP-seq data from 12 transcription factors and 7 histone modifications from mouse embryonic stem

cells they constructed models using support vector regression to investigate how transcription factors and histone modifications interact to predict RNA-seq expression data and the redundancy within and between the transcription factor and histone modification datasets. The level of marker binding was established by taking an area +/- 4kb from the transcription start and end sites of refSeq genes and splitting these regions into 100bp bins, resulting in a total of 160bins for each gene. The coverage of each nucleotide for the transcription factor or histone modification was averaged over the 100bp section and this value was used as the signal for that bin. A signal profile was the calculated as the average signal for the transcription factor or histone modification for each of the bins over all the genes. These signal profiles were then normalised to a total value of 1 to make profiles for the different datasets comparable. Support vector regression was used to predict expression levels for each gene from the transcription factor and histone modification binding signals. The support vector regression was trained and tested multiple times of different testing and training subsets of the data and the average of the resulting Pearson correlation coefficients were used as the correlation between the support vector regression-predicted expression levels and the expression data from RNA-seq or microarray experiments. Unsurprisingly it was found that transcription factor binding is predictive of expression levels. It was also found, using principle component analysis of the bins for the transcription factors, that the influence of the transcription factors extends only as far as 2kb from the transcription start or end sites. Past this range, the contribution to expression of the transcription factor was minimal. Histone modifications were also found to be predictive of transcription, though marginally less so than for the transcription factors. The histone-modification data was found to contribute to expression across the whole of the region investigate though, in contrast to the decreasing contribution at range of the transcription factors. Most interestingly in this study, it was found that transcription factors alone, histone modifications alone or histone modifications and transcription factors combined were roughly equivalent for predicting expression, and the authors go on to highlight the statistical redundancy between the histone modification and transcription factor data.[75] A potential problem with this method of generating the binding signals for the datasets the focus is has on the transcription start and end sites. Histone modifications have different profiles across the gene; some having peaks at the transcription start site that tails off through the gene, while others are more often found as broad peaks that span the majority of the gene-body. Using just an 8000 base region that is centered on the transcription start and end sites is likely to miss the nuances of histone modification presence, especially in cases where the majority of the binding is found in the gene-body.

More recently, Cheng *et al.* have applied the method they previously developed, as de-

scribed above, to a much larger set of data from the ENCODE project.[76] They apply this previous established method to a set of over 400 binding profiles for over 120 transcription factors and histone modifications in "many different cell lines".[76] They use a set of 267 expression profiles for RNA samples from 12 different cell lines. Investigations on this scale are now only just becoming possible with the release of the ENCODE data. The Random Forest machine learning algorithm was used in this investigation for models with multiple predictors, while Support Vector Regression was still used for models with a single predictor. Their aim was to investigate the difference in transcription factor signals at transcription start sites and the contribution of different types of transcription factors to expression. With this vast amount of data from many cell lines, they were also able to show that differential binding of transcription factors is able to explain the differential expression of genes. When investigating the effect of different types of transcription factor, it was found that the presence of non-specific transcription factors was the most predictive of expression, followed by sequence-specific transcription factors and Chromatin restructuring proteins. This is an unsurprising result, as the general transcription factors are so essential for the association of the transcriptional machinery, but goes some way to validate that the method they are using is biologically relevant. The redundancy between the histone modification and transcription factor data was again show here by the comparison of models using the histone modification data, transcription factor data and a combination of the two. A 8% difference in the amount of variance explained when the histone modification data was added to the transcription factor model and a 13% difference in the amount of variance explained was found when the transcription factor data was added to the histone modification model. This shows that while the two difference sets of data are largely statistically redundant, there extra information contained in each that the other lacks.

1.2.7 Importance of understanding genetic regulation

The understanding of genetic regulation is vital for the full understanding of stem cells, cell fate, cancer, many diseases and how cells function. The advances in sequencing technologies have opened up a vast range of new potential studies with the genome-wide identification of transcription factor binding and histone modification presence. Being able to measure variables across the whole genomes means that the subtle interplay between transcription factors, histone modifications and DNA accessibility can now be investigated. The huge volume of data being produced by the new sequencing technologies is complicated. New analytical techniques are being developed to address the issues associated with the volume of data and

the noise that it potentially contains. New statistical methods are also being developed to help understand what the data shows as well as to combine and integrate the wide range of data together.

Investigations into the way in which transcription is regulated, what transcription factors and histone modifications are found in transcribed regions, and the vast and complex network of control that the cell exerts over transcription are now underway. By understanding fully the mechanism that controls transcription the potential for stem cells in treating disease can be unlocked and new, potentially individual-specific, treatments for cancer can be developed.

Chapter 2

Genetic Regulation in Mouse Macrophage Cells

2.1 Introduction

Macrophage Activation and Function

Macrophage cells are generally classified as either classically-activated/M1 macrophages or alternatively-activated/M2 macrophages. More recently macrophages have been described as having three types, with a continuous variation between the types, rather than there being strict differences between the types.[77] Macrophages have three main roles: immune defence, removing damaged tissue and regulating inflammation. Whilst their functions are well defined, macrophages are difficult to classify as all types of macrophage can transition between each of these functions.

Classically-activated macrophages are produced during cell-mediated immune response and also produced in the short-term by stress. They comprise about 90% of the macrophage population in humans.[78] These macrophages are responsible for phagocytosing (engulfing and destroying) bacteria and cell fragments from apoptosis. Classically-activated macrophages are recruited from monocytes circulating in the blood under stimulus from injury, infection or antigen-specific immune cells. Interferon-γ (INFγ) and Tissue Necrosis Factor (TNF) are needed to induce classically-activated Macrophages.[79;80] In the innate immune response, INFγ is produced by natural killer cells as a response to infection or stress. This is only a short-term effect though and the macrophage population is not maintained.[81] T-Helper 1 cells from an adaptive immune response prime macrophage cells to secrete pro-inflammatory cytokines and superoxide anions to help defend against the threat and to increase the pathogen-killing power of the macrophages. To maintain a population of macrophages, signals from an adaptive immune response are needed, such as the antigen-specific response of T-H 1 helper cell and antigen presenting cells. Although there are alternative signals that can replace INFγ and TNF, these two signals together are generally accepted to be the initiators of classically-activated macrophages.[77]

Alternatively-activated macrophages can be split roughly into two populations, the functions of which can overlap. Regulatory-macrophages arise at the end of a period of inflammation; their role is to limit inflammation and reduce the immune response as it is no longer required.[82] Unlike classically-activated macrophages, there doesn't appear to be a single molecule, or pair of molecules, that is responsible for the development of alternatively-activated macrophages, though extracellular-signal-related kinase has been strongly implicated as a mediator.[83] Regulatory macrophages need multiple signals to induce their development, although they do not have a primary pair of molecules.[84] Regulatory macrophages release IL-10, a cytokine with

immunosupressive properties.[85] Wound-healing macrophages, able to be induced by treatment of monocytes with IL-4, are similar to regulatory macrophages but also possess the ability to secrete components of the extracellular matrix.[84] The main purposes of wound-healing macrophages are to to clear up loose, and potentially disruptive cell fragments from the site of a wound, secrete parts of the extracellular matrix and be ready to respond to infection. The function of a macrophage, once differentiated, is not static but can be altered depending on the signals received. This plasticity is an important part of the immune response, allowing the body to react to changes in conditions during infection or would-healing.[84]

2.1.1 Important Transcription Factors

Pu1 (or Spi1) is a cell-fate-determining transcription factor; active genes in macrophage cells are almost always bound by Pu1, especially those with the H3K4Me1 histone modification.[45;86] It is thought that the binding of Pu1 recruits chromatin remodelling factors to change histone modifications on nucleosomes surrounding the binding site.[87] The recruitment of these chromatin remodelling factors opens up the DNA, allowing other factors to bind that would otherwise be unable to bind when the binding sites are within a nucleosome.[88] Although Pu1 is involved in promoting transcriptional activity, its main function is to open up sections of the DNA to allow other transcription factors to bind. Having a single transcription factor act as the enhancer for the majority of genes allows lineage-dependent control of which genes are able to be activated.[89] While Pu1 binding allows regions of the genome to become activated, it does not automatically mean that the transcription apparatus is recruited to the genes or that the cell needs the genes transcribed. Two important transcription factors that control whether a gene in these enhancer regions is active or not are the p65 sub-unit of NFκB (henceforth referred to as p65) and B-Cell Lymphoma protein 6 (Bcl6).

Bcl6, despite being primarily known for the involvement it has in B-Cell differentiation and lymphomas,[90] is found in many cell types and is essential for regulating expression in macrophages.[91] It has been found to maintain a low background level of transcription of genes that are induced by cellular exposure to lipopolysaccharides (LPS) but also limits the maximum rate at which these genes can be transcribed.[92] Exposure of macrophages to LPS has been shown to cause a dramatic (>90%) drop in the binding of Bcl6 to its target genes.[93] The majority of Bcl6 binding occurs at distal sites, often occurring in Pu1 activated regions.[94]

p65 is a well known and studied inflammatory transcription factor which is only able to bind where the DNA is not in a nucleosome.[88] When a cell encounters a signal from Toll-Like

Receptors (TLR) or Tissue Necrosis Factor Receptors (TNFR), p65 is able to bind to its target genes; these are normally inflammatory genes and are often located within Pu1 enhancer regions.[94] The activation of genes by p65 often involves direct or indirect expellation of pre-bound repressors, such as Bcl6, from their binding sites.[92;95]

CCAAT/enhancer-binding protein (CEBP) α and β are members of the bZip family of transcription factors. CEBPα is well-known essential transcription factor for the differentiation of monocytes to macrophage cells when activated by CSF1R.[96] CEBPα binds to the promoter and distal enhancer of Pu1,[97;98] its own promoter,[99] and genes essential for expression of macrophage specific genes and chromatin remodelling.[100] CEBPα initiates the start of differentiation through the expression of Pu1, as well as by forming a feed-forward loop with itself, committing to the act of differentiating. When expressed above a basal level it also inhibits proliferation in cells.[101] Due to this anti-proliferation ability, mutations in CEBPα lead to acute myeloid leukemia (AML)-like transformation in the cells (an early stage in the development of AML).[101] When CEBPα is activated by cAMP-responsive element-binding protein (CREB) it regulates alternatively-activated macrophage-associated genes.[102] Induction of CEBPβ by CREB is only required for genes associated with alternatively-activated macrophages; CEBPβ is involved in the activation of classically-activated macrophages though it is thought that its expression is activated by a different mechanism than CREB binding.[103] Macrophage activation is essential for the production of an inflammatory response to lipopolysaccharides (LPS) molecules present on the surface of gram negative bacteria. Both CEBPα and CEBPβ can, when a macrophage encounters LPS, activate cytokine production to defend against the offending bacterial cell. Experiments have shown that cells lacking CEBPβ or CEBPα are still able to produce these cytokines, though cells with only CEBPα had a lower response to the LPS stimulation than those cells with only CEBPβ.[103] This shows that CEBPα can compensate for the loss of CEBPβ in some circumstances,[103] as it has also been shown that CEBPβ can compensate for the less of CEBPα,[104] though in both cases the effect is less than optimal. Overall, despite their perceived functional redundancy, CEBPα is more involved in the differentiation and maintenance of macrophages while CEBPβ is more involved in immune responses and cytokine production, although they assist one another in fulfilling these roles.

2.1.2 Aims & Objectives

The aims for this chapter are to investigate the relationship between the histone markers (H3K4Me3 and H3K4Me1), transcription factors (Pu1, p65, BCL6, CEBPα and CEBPβ), RNA polymerase II binding, and DNase I hypersensitivity. The objective is to create a simple model that will use the transcription factors, histone modifications and DNase I hypersensitivity to predict the level of RNA polymerase II binding. RNA polymerase II binding is used here as a metric for the level of transcription. The level of RNA polymerase II binding is generally indicative of the level of protein production; RNA Polymerase II binding has been shown to correlate with expression data.[74;105] A simple model for RNA polymerase II binding will help lead to a *mechanistic hypothesis* for the activation of genes. By investigating which of these factors is the most predictive, and how the different factors interact in the models, it will be possible to investigate further into the control of transcription.

2.2 Methods

2.2.1 Data sets

The data used were all from bone marrow derived-mouse macrophage cells, grown in the same conditions. All of the datasets were for unstressed cell (i.e. not treated with lipopolysaccharides). Due to the differences in macrophages throughout the body only macrophages from one tissue-source of macrophages were used. Three experiment sets were used, accessed using the NCBI GEO database[106] or provided by the Bonifer Lab. The datasets used are detailed in table 2.1.

Epigenetic Marker	GEO Accession	Reference
RNA Polymerase II	-	Unpublished Data
H3K4Me3	-	Unpublished Data
CEBP α	-	Unpublished Data
CEBP β	-	Unpublished Data
DNase I Hypersensitivity	GSE26550	Leddin et al.[107]
BCL6	GSE16723	Barish et al.[92]
p65	GSE16723	Barish et al.[92]
IgG	GSE16723	Barish et al.[92]
H3K4Me3	GSE21512	Heinz et al.[86]
Pu1	GSE21512	Heinz et al.[86]

Table 2.1: The datasets used for the epigenetic analysis of mouse macrophage cells

2.2.2 Enrichment Calculations

Read counts were used instead of using complicated peak-finding algorithms and the normalisation calculations. There is currently no normalisation calculation available which does not have some significant problem.[108;109] Instead of using the peak-finding algorithms, a much simpler enrichment calculation was used. The enrichment calculations are normalised within the dataset using a background noise factor directly from the data. This means that the resulting values are comparable between datasets as well as within datasets and removes the need for normalisation between the datasets, which inevitably involves the loss of information. The conservative nature of the background factor should also reduce the number of false positive results. Three different enrichment values were calculated for every gene in each dataset. Firstly, an enrichment value was calculated for the region between the transcription start and end sites (the Gene Enrichment) as defined by the NCBI refSeq entries for the gene.[110] Secondly an enrichment value was calculated for two 2000 base sections, one that spans 2000 bases upstream of the transcription start site and one that starts at the transcription end site

and extends 2000 bases downstream (the Flanking Enrichment). Lastly, an enrichment value was calculated for a region starting 2000 bases upstream of the transcription start site and terminating 2000 bases beyond the transcription end site, thus encompassing the flanking region and the gene in a single enrichment calculation (figure 2.1). 2000 base flanking regions were used as a definition of the flanking as this is the region where it is most likely that the transcription factors will bind.[111]

Figure 2.1: The three regions used for the enrichment calculation. Firstly the number of reads were calculated for the gene, secondly for two 2000 base flanking regions either side of the gene and thirdly for a region that spans the gene and both flanking regions.

$$E = \frac{R_{Region}}{\overline{BG} \times l} \tag{2.1}$$

The enrichment, E, is calculated by dividing the number of reads in the current region, R_{Region}, by the largest average background reads per base, $\overline{BG}$ (described below), multiplied by the length of the current region, l (equation 2.1).

Background Calculation

$$\overline{BG} = \frac{max(R_{Genome}, R_{Chromosome}, R_{Local})}{Bases} \tag{2.2}$$

The background average, $\overline{BG}$, is the biggest average number of reads per base from a selection of potential background areas divided by the number of bases for the current gene (*Bases*, equation 2.2). The areas used for the calculation of the background were the whole of the genome (R_{Genome}), the chromosome on which the gene is located ($R_{Chromosome}$) and a more local region comprising the two 10kbp regions directly surrounding the gene/flanking region (R_{Local}). This background metric was motivated by use of a similar background term in the MACS algorithm for identifying local bias.[71] MACS uses the maximum of a Poisson-distribution-based background, a 2000bp, or a 5000bp window centered on a peak to determine a p-value to filter out false-positive peaks.

2.2.3 Linear Models

Linear models were created using the RNA polymerase II enrichment data as the observation/response variable and the other factors (CEBPα, CEBPβ, DNase I hypersensitivity, BCL6,

p65, IgG, H3K4Me3, H3K4Me1, Pu1) as predictors. The purpose of the linear models was to fit and optimise the combination of the predictors to the RNA polymerase II data. With this platform is is possible to then re-run the models with predictors removed and observe the improvement or degradation in the combination of predictors ability to model RNA polymerase II binding. Both forward and backward stepwise elimination methods were used to build the linear models and both methods resulted in the same optimal regression model in each case.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots \beta_p x_{ip}, \qquad i = 1, \ldots, n, \tag{2.3}$$

Equation 2.3 is the general equation for a linear model, where y_i is the response variable for gene i, x are the regressors (in this case, the enrichment values for the datasets), β_0 is the intercept term and β are the regression coefficients. Linear models were generated using the "lm" function of the R statistical computing package.[112] The log of the enrichment was used for the linear models for better scaling and to ensure a linear relationship. Any enrichment values for which the RNA polymerase enrichment was 0 were removed before the log was taken. Any enrichments of the predictors which had a value of 0 were increased by 0.0001 to allow them to be logged but to have a minimal effect on their value relative to the rest of the data points.

The resulting models were assessed based on their coefficient of determination (R^2) and Bayesian Information Criterion (BIC) score as described in equation 2.4.

$$-2 \cdot \ln p(x|k) \approx BIC = -2 \cdot \ln L + k \ln(n) \tag{2.4}$$

Where x is the observed data, n is the number of data points, k the number of regressors, $p(x|k)$ the probability of the observed data given the number of regressors and L the maximised likelihood function of the model. The BIC value of the model gives an indication of how efficient the model is accounting for the complexity of the model relative to the number of regressors that are used. When comparing two regression models, the one with the lower BIC is preferable.

2.2.4 Outlier and Inlier analysis

The genes which the linear model predicted the enrichment values for RNA polymerase II well and poorly were investigated further. The aim of this was to try and establish whether there were any type or function of genes, such as response effector or differentiation genes, for which these simple models, using only a few important transcription factors and histone modifications, were predicted particularly well or poorly.

For the optimal model and the transcription factor-containing mode for the gene and flanking regions, the logged enrichment values of 5% of the outlying residuals (approximately 1200 highest and lowest outliers) and the residuals closest to 0 (5% of inliers) were analysed for patterns in their Gene Ontology (GO) terms.[113] Hypergeometric tests were used to look for over-enriched GO terms in the outlying and inlying genes of the transcription-factors containing model and the optimal model of the gene and flanking logged enrichments. Hypergeometric testing was performed on the outliers and the inliers to assess any over-representation in GO terms within these sets of genes. This was performed using the GOstats package provided by Bioconductor[114] in the R statistical program.[112]

The inlying and outlying genes were also compared to genes that are known to be important to macrophage development taken from Ravasi *et al.*[115] (table 2.2).

Gene	Function	Gene	Function
ATF3	Regulation of inflammatory response	MYC	Determination of macrophage type
BCL3	Transcriptional co-activator	NCOA3	Macrophage differentiation
BCL6	Transcriptional repressor	NCOR1	Inflammatory Regulation
CEBPA	Role in hematopiesis, immune response	NCOR2	Inflammatory Regulation
CEBPB	Regulates genes for immune/inflammatory response	NFE2L2	Oxidative Stress Resistance
CEBPD	Regulates genes for immune/inflammatory response	NFKB2	Macrophage differentiation and immune response
CITED2	Transcriptional repressor	NFKB1	Macrophage differentiation and immune response
CREB1	Response to hormonal stimulation	NFKBIA	Macrophage differentiation and immune response
DDIT3	IL-6 regulation	NR1H2	Inflammatory gene expression
EGR1	Macrophage differentiation, adhesion, Phagocytosis	NR1H3	Inflammatory gene expression
EGR2	Macrophage differentiation, adhesion, Phagocytosis	NR3C1	Inflammatory Regulation
ELF1	Macrophage-specific gene expression	p65 (RELA)	Inflammatory Regulation
EP300	Endotoxin-stimulated gene expression	POU2F2	Immune Response
ETS2	Immune Response	PPARG	Determination of macrophage type
FOS	Macrophage differentiation	Pu1 (SPI1)	Macrophage differentiation and immune response
FOSL1	Macrophage differentiation	RARA	Macrophage differentiation
FOXO3A	Apoptosis	RB1	Macrophage differentiation
GFI1	Endotoxin Tolerance	RELB	Immune Response
HDAC1	Growth, immune response	RUNX1	Regulated growth and survival
HDAC3	Growth, immune response	RXRA	Macrophage differentiation
HIF1A	Tumor Response	RXRB	Macrophage differentiation
HSF1	Immune Response	SP1	Immune Response
IRF1	Immune Response	SP3	Immune Response
IRF2	Immune Response	STAT1	Tumor Response
IRF3	Immune Response	STAT3	Immune Response
IRF5	Immune Response	STAT5A	Immune Response
IRF7	Immune Response	STAT5B	Immune Response
JUN	Immune Response	STAT6	Immune Response
JUNB	Immune Response	TFE3	Macrophage differentiation
JUND	Immune Response	TFEC	Macrophage-specific gene regulation
KLF4	Determination of macrophage type	TRIM28	Macrophage differentiation
MAFB	Macrophage differentiation	YY1	Inflammatory response
MITF	Immune Response		

Table 2.2: A list of genes known to be important to the development or function of mouse macrophage cells taken from Ravasi *et al.*[115]

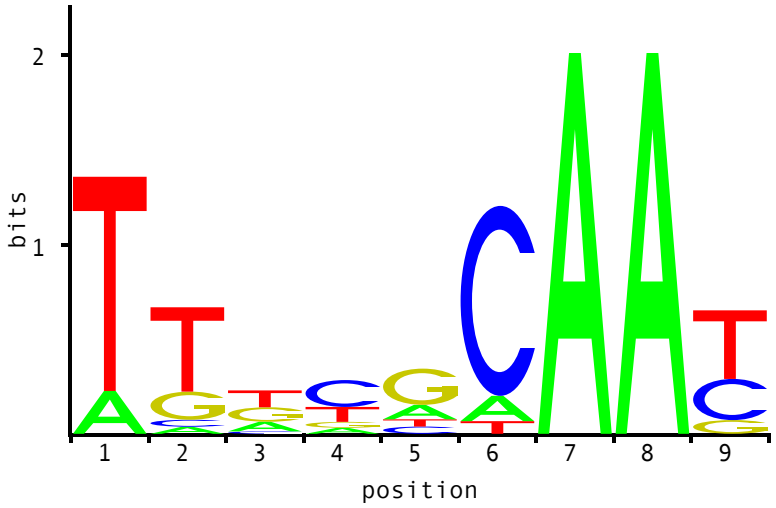
2.2.5 Drop Analysis

The effect of the transcription factors on the linear model results was investigated by comparing the best model containing transcription factors to a similar model with one of the transcription factors removed. This was done for CEBPα, p65 and Pu1, of pairs of the transcription factors, and removing all three transcription factors from the model. The gene and flanking logged enrichment model containing the transcription factors was used as the base model from which the transcription factor was removed. The two linear models were then compared to identify the genes that were affected the most and least by the removal of the transcription factor. These genes were then investigated as to whether they were known or likely targets for the transcription factor.

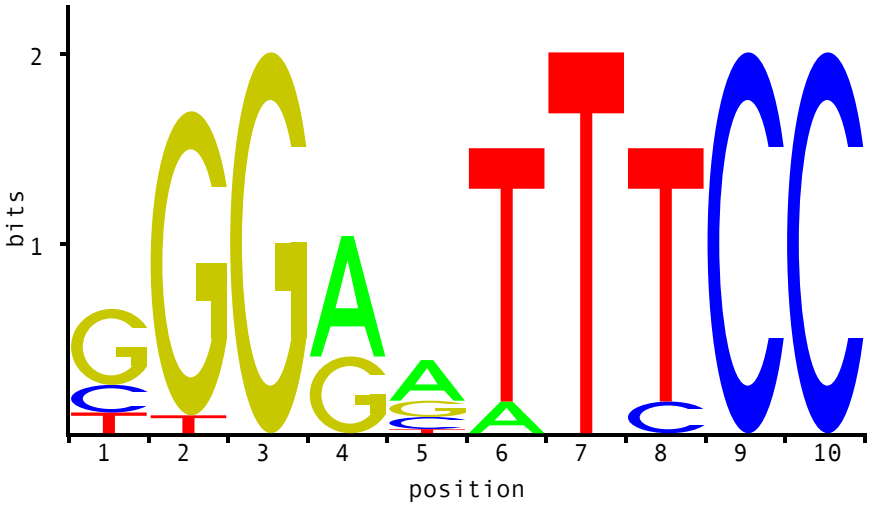
Predicting Transcription Factor Binding Sites

Relationships between most/least affected genes and transcription factors was investigated through transcription factor binding site prediction and comparison to conserved binding sites. MSigDB is a database of annotated gene sets.[116] The C3 set of MSigDB (a set of cis-regulatory motifs that are conserved between Human, Rat, Mouse and Dog genomes) was used to compare genes that were affected by the model minus the transcription factor(s) to genes that are known to be bound by these transcription factors.

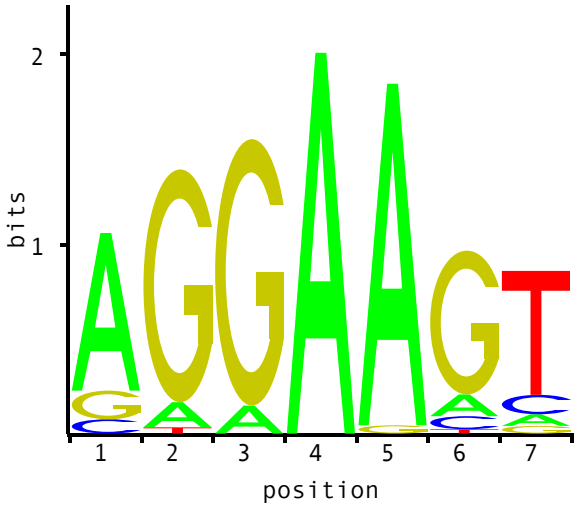
MOODS[61] was used to predict the binding sites for the transcription factors across the whole of the genome using the consensus sequences from JASPAR.[59] MOODS uses position-weight matrices provided from JASPAR to find the most significant sub-matrix of the motif. This sub-matrix represents the most likely core of the binding motif. MOODS then searches across the genome to identify any location where this sub-matrix is found. The identified sub-matrix is then extended and scored using log-odds against a background distribution.[59] The distribution of the p-values for any predicted binding sites located between the start and end of the gene/flanking region were then compared to the genes that were most and least affected by the removal of the transcription factor(s) from the model and 2000 randomly selected genes. The consensus sequences used are shown in figure 2.2.



(a) Consensus sequence for the DNA binding site of CEBPα from JASPAR.[59]



(b) Consensus sequence for the DNA binding site of p65 from JASPAR.[59]



(c) Consensus sequence for the DNA binding site of Pu1 from JASPAR.[59]

Figure 2.2: Binding motifs from JASPAR[59] used in the prediction of binding sites.

2.3 Results

2.3.1 Linear Models

To investigate and develop a mechanistic hypothesis regarding the occurrence of RNA polymerase binding, linear models were used to fit the logged enrichment data for the histone modification, transcription factor and DNase I hypersensitivity to the RNA polymerase II data. RNA polymerase II binding was used as a measure of transcription on the assumption that the more RNA polymerase II that is bound, the more RNA will be produced. Data from RNA polymerase II ChIP-Seq experiments has been shown previously to correlate with expression data.[74;105] Although not always equal, the amount of RNA that is produced is normally a good indication of the amount of protein produced. The predictors (the transcription factors, histone modifications and DNase I hypersensitivity) were eliminated using a backward elimination method, whereby the most non-significant (lowest p-value) predictor was removed from the model. These models were checked by building a forward addition method, where predictors with the most positive influence were added incrementally to an empty model; this gave the same result. p-values were not used to evaluate the quality of the model as with such a large dataset a small p-value can still result in a large number of badly predicted points. Instead a combination of the adjusted-R^2, f-statistic and Bayesian Information Criterion (BIC) were used. The adjusted-R^2 shows the amount of variance in the RNA polymerase II data that can be explained by the model. The f-statistic and BIC give an indication as to whether the model is losing (or gaining) information relative to the number of predictors in the model. The linear models were designed in order to find the simplest model that could explain the majority of the variance in the RNA polymerase II data while using the fewest number of predictors. A simple model allows us to eliminate the genes which are easily predicted as transcribed or not. The genes for which the model does not work can be highlighted as genes which may use other mechanisms which the simple model does not account for.

The optimal model for the logged enrichments for the gene and flanking data was one comprising of just DNase I hypersensitivity, H3K4Me1 and H3K4Me3 (table 2.3). This model had an adjusted-R^2 of 0.7832 and a BIC of 50564. The adjusted-R^2 was lower than a model that contained H3K4Me3, H3K4ME1, DNase I hypersensitivity, p65, Pu1 and CEBPα, which had an adjusted-R^2 of 0.7911, but had a larger BIC. This indicates that the information contained in the majority of the datasets is not necessary for predicting the binding of RNA polymerase II. The increased BIC in the more complex model shows that the smaller model still retains the information needed to predict the binding. It is interesting that the optimal model for the gene and

Model	Adjusted-R^2	p-value	f-statistic	Predictors withheld	BIC
PolII $\sim$ CEBPα + CEBPβ + DNase + H3K4Me3 + BCL6 + IgG + p65 + H3K4Me1 + Pu1	0.7917	2.2x10^{-16}	9655 on 9 and 22856 DF		51564
PolII $\sim$ CEBPα + DNase + H3K4Me3 + BCL6 + IgG + p65 + H3K4Me1 + Pu1	0.7916	2.2x10^{-16}	1.086x10^{-4} on 8 and 22857 DF	-CEBPβ	50908
PolII $\sim$ CEBPα + DNase + H3K4Me3 + IgG + p65 + H3K4Me1 + Pu1	0.7913	2.2x10^{-16}	1.239x10^{-4} on 7 and 22858 DF	-CEBPβ, BCL6	51101
PolII $\sim$ CEBPα + DNase + H3K4Me3 + p65 + H3K4Me1 + Pu1	0.7911	2.2x10^{-16}	1.443x10^{-4} on 6 and 22859 DF	-CEBPβ, BCL6, IgG	51243
PolII $\sim$ CEBPα + DNase + H3K4Me3 + H3K4Me1 + Pu1	0.7895	2.2x10^{-16}	1.716x10^{-4} on 5 and 22860 DF	-CEBPβ, BCL6, IgG, p65	51165
PolII $\sim$ CEBPα + DNase + H3K4Me3 + H3K4Me1	0.7864	2.2x10^{-16}	2.120x10^{-4} on 4 and 22861 DF	-CEBPβ, BCL6, IgG, p65, Pu1	50738
PolII $\sim$ DNase + H3K4Me3 + H3K4Me1 + Pu1	0.7876	2.2x10^{-16}	2.120x10^{-4} on 4 and 22861 DF	-CEBPβ, BCL6, IgG, p65, CEBPα, Pu1	50717
PolII $\sim$ DNase + H3K4Me3 + H3K4Me1 + p65	0.7838	2.2x10^{-16}	2.072x10^{-4} on 4 and 22861 DF	-CEBPβ, BCL6, IgG, CEBPα, Pu1	50708
PolII $\sim$ DNase + H3K4Me3 + H3K4Me1	0.7832	2.2x10^{-16}	2.753x10^{-4} on 3 and 22862 DF	No Transcription factors	50701
PolII $\sim$ CEBPα + Pu1 + p65	0.4797	2.2x10^{-16}	7029 on 3 and 22862 DF	Transcription factors from the second best model only	71575

Table 2.3: Analysis of the Linear models containing different combinations of the predictors for the flanking regions and gene region logged enrichments.

flanking data doesn't employ any of the transcription factor data as a predictor. More complex models, such as ones containing CEBPα, DNase, H3K4ME3, H3K4Me1, p65 and Pu1, have a slightly higher correlation with the RNA polymerase II logged enrichments than the simpler model but result in a higher BIC. Comparing the models that contain transcription factors, the optimal model (with CEBPα, with p65, with Pu1 or with all three of these transcription factors) shows that their addition does not lead to a significant improvement in the correlation of the model compared to the simpler model. The models with p65, Pu1 or CEBPα have adjusted-R^2 of 0.7838, 0.7876 and 0.7864 respectively. Compared to the adjusted-R^2 of 0.7832 in the simple model, there isn't a large enough increase in the correlation to warrant choosing a complex model over a simpler model with fewer predictors, a comparable adjusted-R^2 and a smaller BIC.

Model	adjusted-R^2	p-value	f-statistic	Predictors withheld	BIC
PolII $\sim$ CEBPα + CEBPβ + DNase + H3K4Me3 + BCL6 + IgG + p65 + H3K4Me1 + Pu1	0.7708	2.2x10^{-16}	8474 on 9 and 22666 DF		51986
PolII $\sim$ CEBPα + DNase + H3K4Me3 + BCL6 + IgG + p65 + H3K4Me1 + Pu1	0.7708	2.2x10^{-16}	9531 on 8 and 22667 DF	-CEBPβ	51405
PolII $\sim$ CEBPα + DNase + H3K4Me3 + IgG + BCL6 + H3K4Me1 + Pu1	0.7706	2.2x10^{-16}	1.088x10^{-4} on 7 and 22668 DF	-CEBPβ, p65	50830
PolII $\sim$ CEBPα + DNase + H3K4Me3 + IgG + H3K4Me1 + Pu1	0.7702	2.2x10^{-16}	1.267x10^{-4} on 6 and 22669 DF	-CEBPβ, BCL6, p65	50682
PolII $\sim$ CEBPα + DNase + H3K4Me3 + H3K4Me1 + Pu1	0.7686	2.2x10^{-16}	1.507x10^{-4} on 5 and 22670 DF	-CEBPβ, BCL6, IgG, p65	50649
PolII $\sim$ DNase + H3K4Me3 + H3K4Me1	0.7626	2.2x10^{-16}	1.850x10^{-4} on 4 and 22671 DF	-CEBPβ, BCL6, IgG, p65, DNase	50647
PolII $\sim$ H3K4Me3 + H3K4Me1 + Pu1	0.7563	2.2x10^{-16}	2.346x10^{-4} on 4 and 22672 DF	-CEBPβ, BCL6, IgG, p65, DNase, CEBPα	50651

Table 2.4: Analysis of the Linear models containing different combinations of the predictors for the Gene region logged enrichments.

The optimal model for the logged enrichments of the gene data is a model comprised DNase I hypersensitivity, H3K4Me3 and H3K4Me1 (table 2.4). This model was close in both adjusted-R^2 and BIC to a model comprising of H3K4Me3, H3K4Me1 and Pu1. The model containing DNase as opposed to Pu1 is superior in that it has a marginally better adjusted-R^2 and a lower BIC compared the Pu1-containing model. It is interesting, but not surprising,

to note that the optimal model for the gene data is the same for the gene and flanking data, though the steps used to get to each of the models differ. This is likely because the gene and flanking data is largely the gene data, so the best models for each are likely to be similar.

Model	adjusted-R^2	p-value	f-statistic	Predictors withheld	BIC
PolII $\sim$ CEBPα + CEBPβ + DNase + H3K4Me3 + BCL6 + IgG + p65 + H3K4Me1 + Pu1	0.7026	2.2x10^{-16}	4161 on 9 and 15844 DF		48690
PolII $\sim$ CEBPα + CEBPβ + DNase + H3K4Me3 + BCL6 + IgG + H3K4Me1 + Pu1	0.7024	2.2x10^{-16}	4678 on 8 and 15845 DF	-p65	45229
PolII $\sim$ CEBPα + CEBPβ + DNase + H3K4Me3 + BCL6 + H3K4Me1 + Pu1	0.7023	2.2x10^{-16}	5343 on 7 and 15846 DF	-p65, IgG	45214
PolII $\sim$ CEBPα + DNase + H3K4Me3 + BCL6 + H3K4Me1 + Pu1	0.7020	2.2x10^{-16}	6225 on 6 and 15847 DF	-p65, IgG, CEBPβ	45209
PolII $\sim$ CEBPα + DNase + H3K4Me3 + BCL6 + Pu1	0.7017	2.2x10^{-16}	7460 on 5 and 15848 DF	-p65, IgG, CEBPβ, H3K4Me1	45203
PolII $\sim$ CEBPα + DNase + H3K4Me3 + Pu1	0.7013	2.2x10^{-16}	9305 on 4 and 15849 DF	-p65, IgG, CEBPβ, H3K4Me1, BCL6	45201
PolII $\sim$ CEBPα + DNase + Pu1	0.6282	2.2x10^{-16}	8930 on 4 and 15850 DF	-p65, IgG, CEBPβ, H3K4Me1, BCL6, H3K4Me3	45205

Table 2.5: Analysis of the Linear models containing different combinations of the predictors for the flanking regions logged enrichment.

The optimal model for the flanking region (table 2.5) was a model containing CEBPα, DNase I hypersensitivity, H3K4Me3 and Pu1 as it had the lowest BIC. It is interesting that the models for all of the datasets contain both DNase I hypersensitivity and the H3K4Me3 histone modification, indicating that it is likely that these two predictors are a good indication of RNA polymerase II binding. H3K4Me3 is known to be linked with the activation of transcription. DNase I hypersensitivity is indicative of the DNA uncoiling from the histones and the tighter chromatin structure, which exposes it to allow RNA polymerase II binding.

Overall, the best result from using the linear models to predict the binding of RNA polymerase II was from the model using both the gene and flanking data (figure 2.3, residuals in figure 2.4). This model had a higher adjusted-R^2 than the models for the gene or flanking data, as well as a lower BIC, indicating a greater amount of information relative to the num-

ber of predictors and a better correlation with the RNA polymerase II binding data. The data which combines the flanking and gene regions together potentially has more information than looking exclusively at the gene or flanking regions. Transcription factors tend to have binding sites withing the flanking/promoter region, while histones markers, DNase I hypersensitivity and RNA polymerase II binding tend to occur between the transcription start and end sites (the gene region).

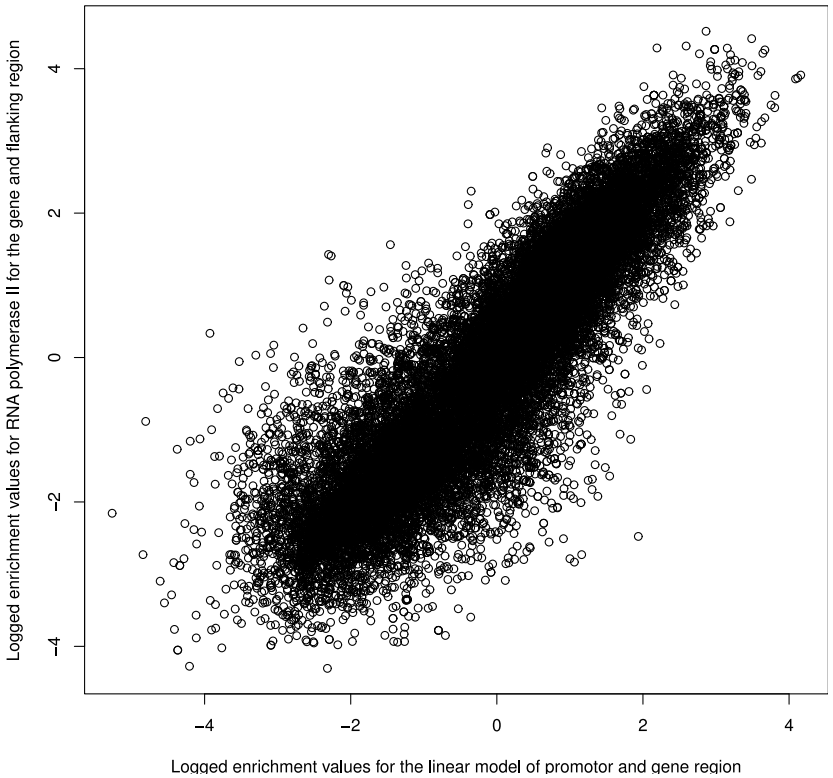


Figure 2.3: The optimal linear model for the logged enrichment of the flanking and gene data. The optimal model predicted RNA polymerase II binding using DNase I hypersensitivity, H3K4Me3 and H3K4Me1 presence. Adjusted-R^2=0.7832. p-value=2.2x10^{-16}. BIC=50564

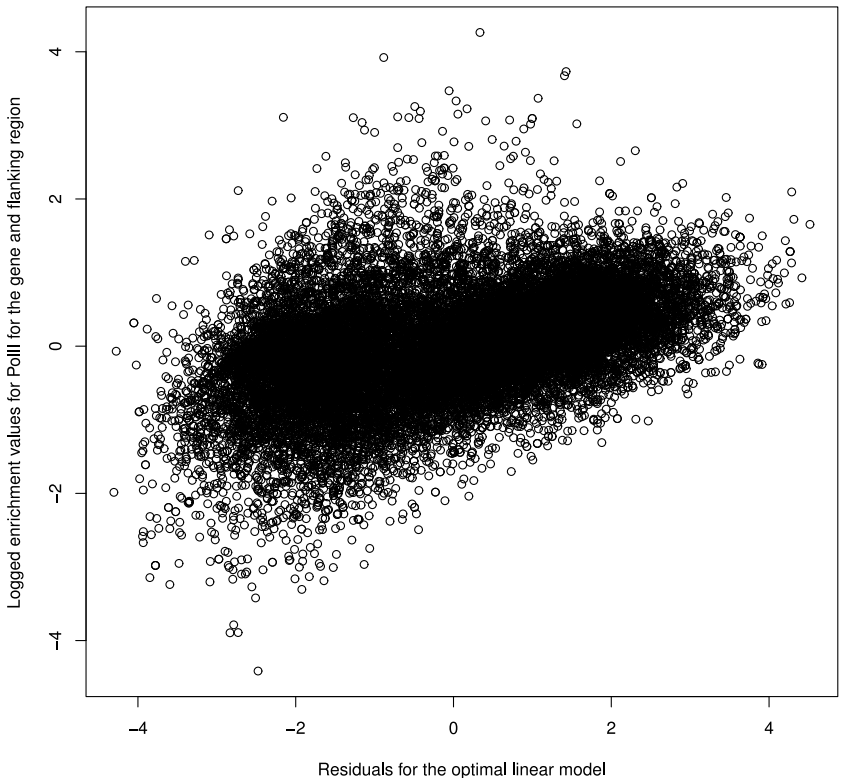


Figure 2.4: The residuals of the linear model predicting the log enrichment values for RNA polymerase II enrichment using H3K4Me3, H3K4Me1, DNase I hypersensitivity, CEBPα, p65 and, Pu1. Adjusted-R^2= 0.7911, p-value=2.2x10^{-16}, f-statistic=1.443x10^{-4}, BIC=50738.

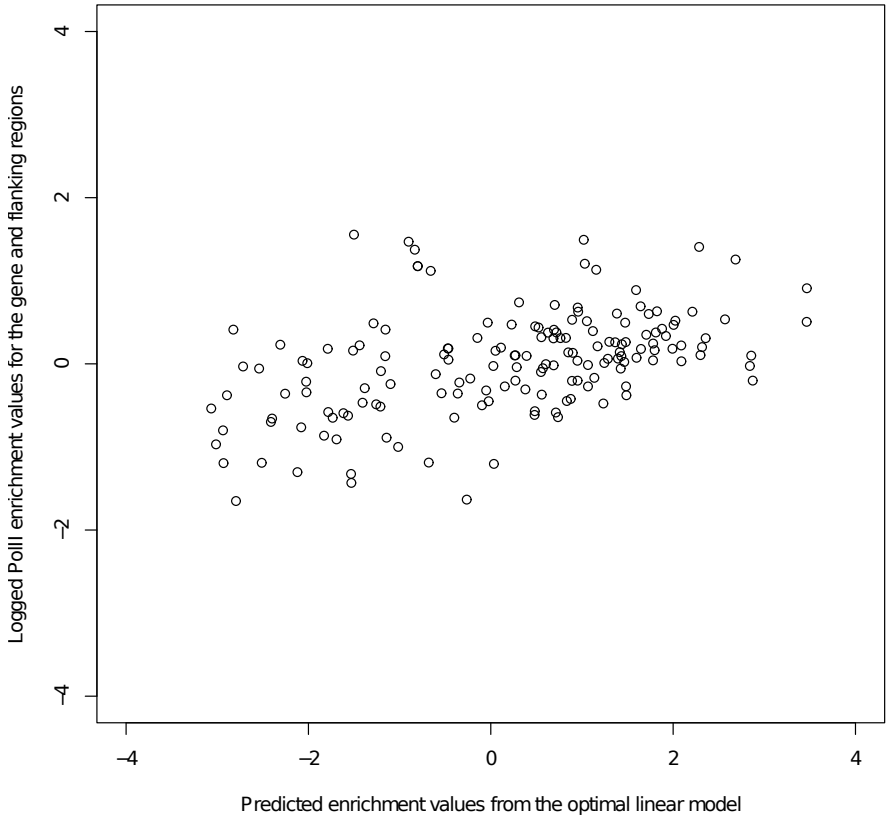


Figure 2.5: The residuals for only the macrophage-specific genes as detailed in table 2.2 for the logged enrichment linear model, using the gene and flanking regions, using H3K4Me3, H3K4Me1, DNaseI hypersensitivity, CEBPα, p65 and Pu1.

Looking at the residuals for the genes that have a specific relation to macrophage cells (table 2.2) it is clear that there is a spread of genes throughout the range of predicted values; the pattern of predicted enrichment values neither stands out as being consistently well predicted or consistently poorly predicted (figure 2.5). The genes that have a positively predicted enrichment value (i.e. are predicted to be expressed) are predicted better than those with a negative predicted enrichment value (i.e. genes predicted to not be expressed), as evidenced by the much tighter clustering on the right of figure 2.5. This is likely due to the amount of noise present in the negative enrichment values; any true signal is likely to be affected more by the background noise in the ChIP-seq data. While it is disappointing that the residuals for the macrophage-specific genes are not more tightly clustered along the 0-line, the point at which the predicted value matches the actual enrichment value of the RNA polymerase II data, the genes are all predicted reasonably well, and the majority lie well within the bulk of the residuals. The reason that the macrophage-related genes are not predicted better is likely due to the models being trained on all of the genes, and not just this subset of genes. Constructing the model using all of of the genes, rather than ones specific to a certain function,

means that the model is optimised to predict the enrichment values of any gene in general. It is likely that if the models were trained on a subset of the genes, such as the macrophage-related genes, it would perform better at predicting these types of genes or genes that had a similar transcription-regulatory mechanism, than genes with different regulatory mechanisms or functions. While it would be interesting to do this, it is often statistically unfeasible to do due to the small numbers of genes in these subsets.

2.3.2 Analysis of the Inliers and Outliers

The 5% of genes with the largest positive or negative residuals (outliers) and the 5% of genes with the smallest residuals (inliers) for the model predicting RNA polymerase II logged enrichment using H3K4Me3, H3K4Me1, DNase I hypersensitivity, CEBPα, p65 and Pu1 were analysed for patterns. The aim of this further analysis was to investigate which genes the model worked poorly for, with the intention of identifying macrophage-specific genes, functions or processes that the model predicted poorly. Being able to identify the genes for which the model *doesn't* work will give us insight into any biological processes or events which are not accounted for by the limited number of predictors used in the model.

Firstly, the genes found in these two sets were compared to a set of genes known to be specifically relevant to macrophages[115] and from genes known or predicted to be bound by the transcription factors in the MSigDB database.[116] There were no genes found in the inliers or outliers present in either of these macrophage-relevant genes.

Hypergeometric testing was used to determine whether there was any over-enrichment of any GO terms, with the intention of highlighting any specific process, function or localisation for which the model was a poor predictor. Table 2.6 shows the results of the hypergeometric testing. For the outlying genes, there were a total of 11 significantly enriched (adjusted p-value < 0.001) GO terms, seven in the molecular function ontology, three in the biological process ontology and one in the cellular component ontology. Four of the seven molecular function enriched GO terms were related to DNA binding and, the other three were peptidase or peptidase inhibitor activity related. Overall, the outliers show no macrophage or immune-specific over-enrichment of GO terms.

For the inlying genes, there were a total of 10 significantly enriched (adjusted p-value < 0.001) genes: one in the molecular function ontology, none in the biological process ontology, and nine in the cellular component ontology. All nine of the over-enriched GO terms were interconnected and revolved around the over-enrichment of the membrane-bound organelle

term. As with the outlying genes, there was no over-representation of any macrophage or immune-specific GO terms.

Outliers			
GO ID	GO Term	Ontology	p-value
GO:0006366	transcription from RNA polymerase II promoter	BP	4.31×10^{-5}
GO:0006357	regulation of transcription from RNA polymerase II promoter	BP	1.01×10^{-4}
GO:0006355	regulation of transcription, DNA-dependent	BP	9.37×10^{-4}
GO:0005576	extracellular region	CC	2.97×10^{-11}
GO:0043565	sequence-specific DNA binding	MF	2.02×10^{-11}
GO:0003700	sequence-specific DNA binding transcription factor activity	MF	1.11×10^{-10}
GO:0001071	nucleic acid binding transcription factor activity	MF	1.19×10^{-10}
GO:0003677	DNA binding	MF	2.01×10^{-6}
GO:0030414	peptidase inhibitor activity	MF	3.49×10^{-4}
GO:0004866	endopeptidase inhibitor activity	MF	6.50×10^{-4}
GO:0008236	serine-type peptidase activity	MF	9.94×10^{-4}
Inliers			
GO ID	GO Term	Ontology	p-value
GO:0005488	binding	MF	6.16×10^{-6}
GO:0005623	cell	CC	1.91×10^{-9}
GO:0044464	cell part	CC	2.92×10^{-9}
GO:0005622	intracellular	CC	7.74×10^{-9}
GO:0044424	intracellular part	CC	5.27×10^{-8}
GO:0043231	intracellular membrane-bounded organelle	CC	1.72×10^{-5}
GO:0043227	membrane-bounded organelle	CC	2.15×10^{-5}
GO:0043229	intracellular organelle	CC	4.48×10^{-5}
GO:0043226	organelle	CC	5.83×10^{-5}
GO:0005737	cytoplasm	CC	1.65×10^{-4}

Table 2.6: The over-represented GO terms from the genes with the 5% largest (outliers) and 5% smallest (inliers) for the linear model predicting the logged enrichment values of RNA polymerase II using H3K4Me3, H3K4Me1, DNase I hypersensitivity, p65, Pu1 and CEBPα logged enrichments as predictors. The GO ontologies are: BP (Biological Process), CC (Cellular Component) and MF (Molecular Function).

Overall, the best-fit and worst-fit genes for the model appear to have no over-enrichment for macrophage-specific genes. This could be due to the number of genes in the analysis masking the location of the macrophage-specific genes under genes that are modelled less well for unknown reasons.

2.3.3 Drop Analysis on linear model

The best model produced was a model that contained no transcription factors. This is interesting as transcription factors are essential for the activation of genes and are ultimately responsible for the modifications to the histones; their absence is therefore noteworthy. The effect of each transcription factor on the model was looked at in more depth. Using the model

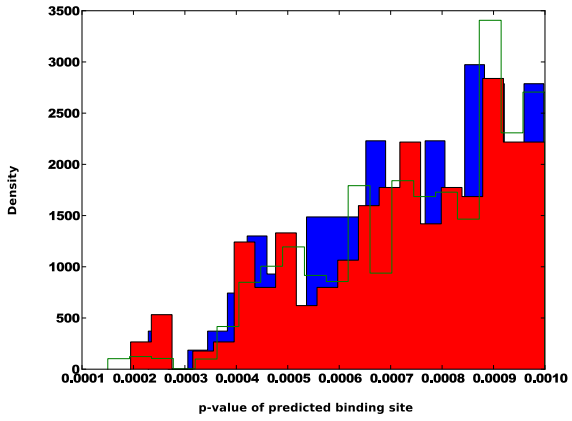
containing H3K4Me3, H3K4Me1, DNase I hypersensitivity, Pu1, p65 and CEBPα as a baseline, each of the transcription factors were then removed in turn. The two models - with and without the transcription factors - were then compared to see which genes were affected the most by the change. For example, CEBPα was removed and this model was compared to a model only containing H3K4Me3, H3K4Me1, DNase I hypersensitivity, Pu1 and p65. The genes most and least affected by this change were then studied further for any patterns in the genes themselves and any potential transcription factor binding sites for the transcription factor that was removed.

When a comparison was made between the genes most affected by the removal of each transcription factor and the genes known to be important for macrophage development and function (table 2.2) there were no matching genes. When compared to a set of genes from MSigDB[116] for which it is known that each transcription factor is bound there were also no matches between the two lists of genes. The genes that were shown to be affected most, and least, by the removal of a transcription factor where then analysed further to see if they had the potential to be targets for the transcription factor in question.

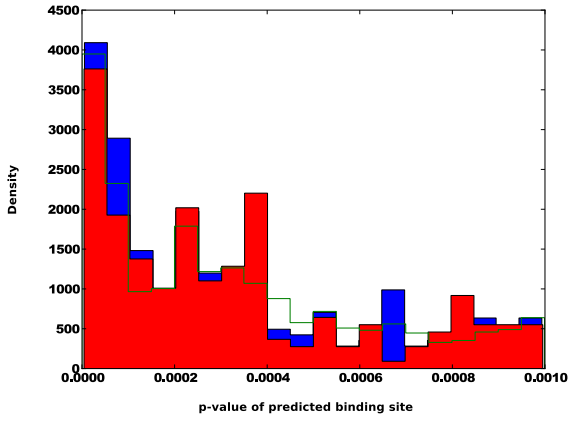
When each of the transcription factors were removed from the model, the distribution of the p-values for the gene binding sites that were affected the most by the change showed very little difference when compared to the distributions for the least affected genes and 2000 random genes (figure 2.6). This observation was repeated when pairs of the transcription factors were dropped (figure 2.7) and when all three transcription factors dropped (figure 2.8) and the distributions of the transcription factors compared.

The very strong similarity between the distributions of the p-values for predicted binding sites for the affected and unaffected genes with the random genes for all the models, regardless of how many transcription factors were removed, indicated that they add no extra information to the model beyond the histone modifications and DNase I hypersensitivity. The transcription factors failing to add any more information beyond the histone modifications could be the result of several conditions. The transcription factors, having bound to the DNA, are what cause the histones to be modified and thus become accessible to RNA polymerase II. This dependency of these modifications on transcription factors could mean that the effect of the transcription factors is being masked by the stronger statistical effect of the histone modifications in the model. Equally, it could be that the transcription factors are only able to bind in the same regions that the histone modifications are in because of the modifications. This again would statistically mask the effect the transcription factors have on the binding of RNA polymerase II. The statistical redundancy of the transcription factors and the histone modifications

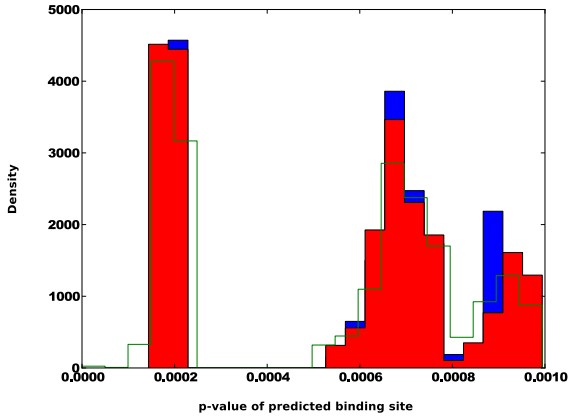
has been noted recently.[117]



(a)



(b)



(c)

Figure 2.6: Comparison between the MOODS[61] predicted transcription factor binding sites the genes that were affected the most (blue) and least when (a) CEBPα, (b) p65, and (c) Pu1 were removed from the best transcription factor-containing linear model. The green line indicates the distribution of the p-values of the binding sites within 2000 randomly selected genes. Concensus sequences for each transcription factor were provided by JASPAR.[59]

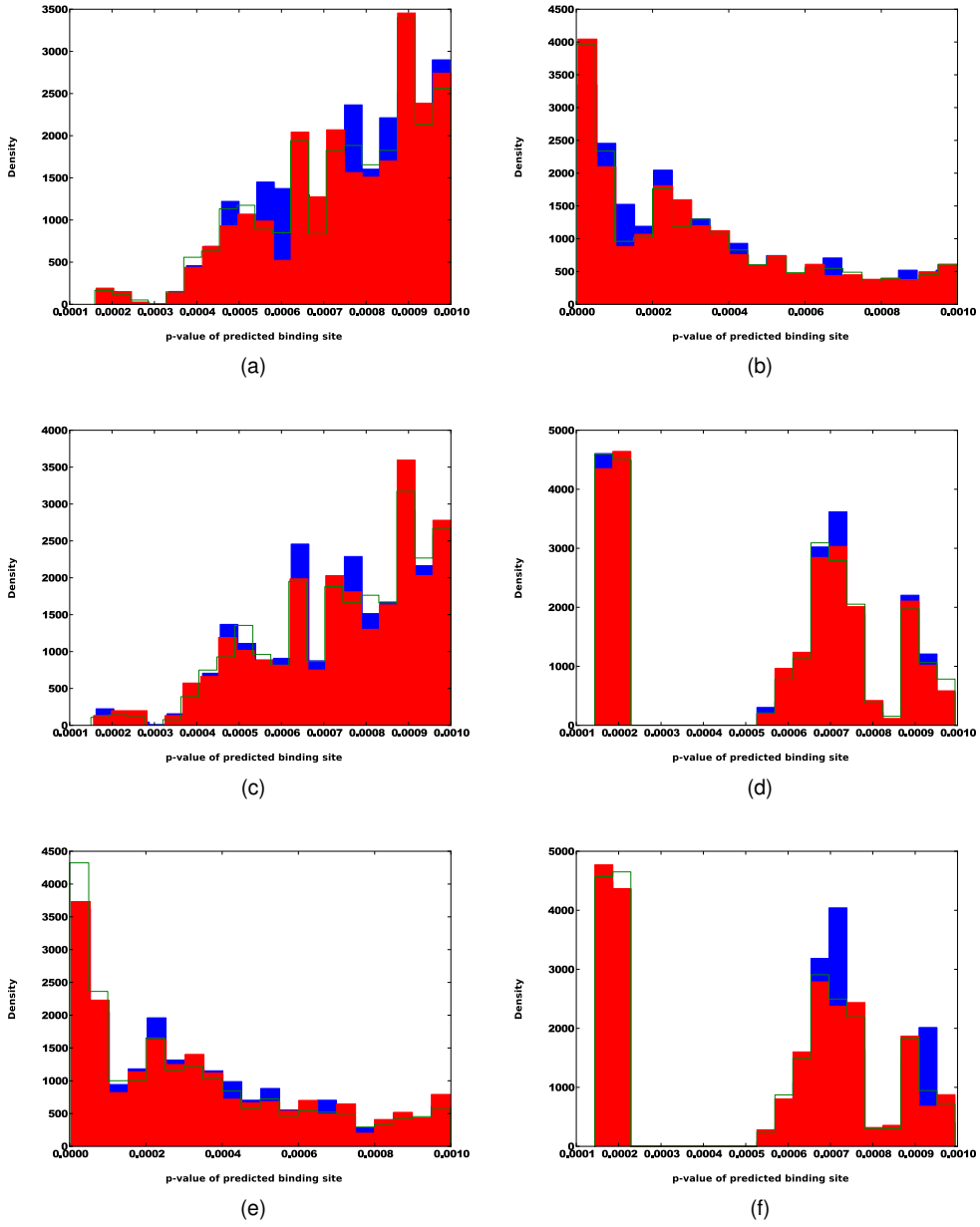
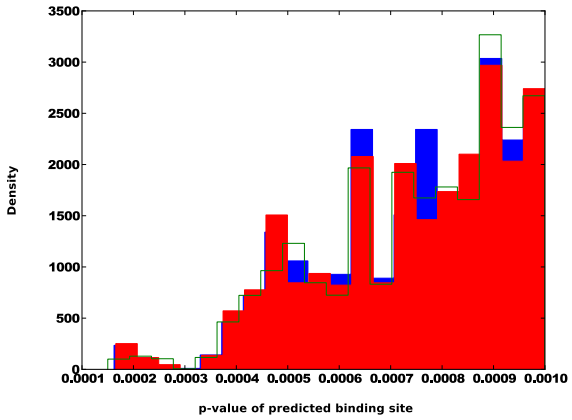
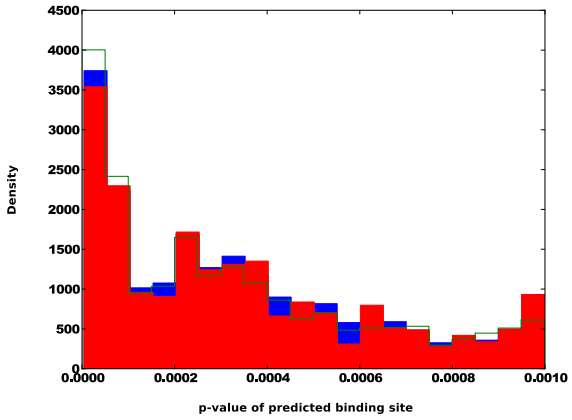


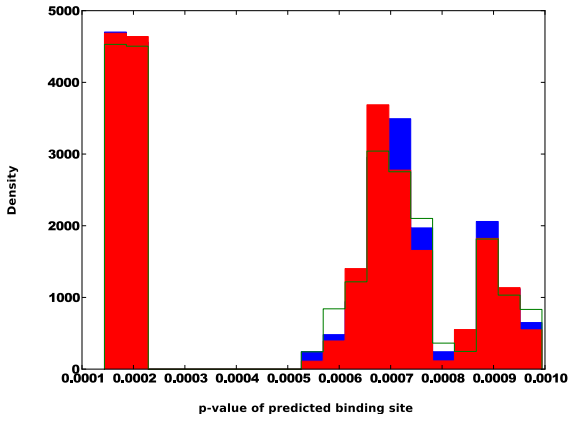
Figure 2.7: Comparison between the MOODS[61] predicted transcription factor binding sites for genes that were affected the most (blue) and least (red) when pairs of transcription factors were removed from the best transcription factor-containing model. The green line indicates the distribution of of the p-values of the binding sites within 2000 randomly selected genes.Concensus sequences for each transcription factor were provided by JASPAR.[59] (a) Distribution of CEBPα binding sites when CEBPα and p65 were removed from the model. (b) Distribution of p65 binding sites when CEBPα and p65 were removed from the model. (c) Distribution of CEBPα binding sites when CEBPα and Pu1 were removed from the model. (d) Distribution of Pu1 binding sites when CEBPα and Pu1 were removed from the model. (e) Distribution of p65 binding sites when Pu1 and p65 were removed from the model. (f) Distribution of Pu1 binding sites when Pu1 and p65 were removed from the model.



(a)



(b)



(c)

Figure 2.8: Comparison between the MOODS[61] predicted transcription factor binding sites for genes that were affected the most (blue) and the least (red) when all three transcription factors were removed from the best transcription factor-containing model. The green line indicates the distribution of the p-values of the binding sites within 2000 randomly selected genes. Consensus sequences for each transcription factor were provided by JASPAR.[59] (a) Distribution of CEBPα binding sites when CEBPα, Pu1 and p65 were removed from the model. (b) Distribution of p65 binding sites when CEBPα, Pu1 and p65 were removed from the model. (c) Distribution of Pu1 binding sites when CEBPα, Pu1 and p65 were removed from the model.

2.3.4 Stratification

With the intent of determining if there is a difference between the genes with a high versus a low enrichment value, the data were split in several ways to investigate the different models for each subset of data. For the analysis, the data were split in three ways (figure 2.9). Firstly, the RNA polymerase II data is clearly bimodal and was divided as such. The genes with high expression were described as everything above -0.7 and the genes with low expression as having a log enrichment value $\leq$ -0.7. Secondly, the data was split into high and low expression genes: high expression genes as those above 0 and low expression as those with a log enrichment value $\leq$ 0. It is important to note that an enrichment value below 0 means that there are fewer reads for the gene in the RNA polymerase II data than in the background metric that was used to calculate the enrichment; this is a good indicator that the gene is hardly expressed or not expressed at all. The final two analyses of these sub-sets of data were similar to the previous ones, the genes with low expression were regarded as those with a RNA polymerase II log enrichment value $\leq$ 0. The genes with a log enrichment value above 0 were split into two subsets, the first were the genes with medium expression (as those with a log enrichment value between 0 and 1.75) and the genes with high expression (those with a log enrichment value above 1.75) for the first of the two analyses. The second of these final two analyses took the genes with medium expression (those with log enrichment scores between 0 and 2) and the genes with high expression (those with log enrichment scores over 2). The first of these final two sub-sets uses the medium-high boundary so that it splits the number of genes that have a log enrichment value above 0 in to two equal sections, whilst the second medium-high boundary is based on splitting the region above 0 in to two by value.

	Limits	Optimal Model	R^2	BIC
High	>-0.7	PolII $\sim$ CEBPα + DNase I HS + H3K4Me3	0.6339	25100
Low	<-0.7	PolII $\sim$ CEBPα + CEBPβ + DNase I HS + H3K4Me3 + p65 + H3K4Me1 + Pu1	0.3040	14004
High	>0	PolII $\sim$ H3K4Me3 + DNase I HS + Pu1	0.6021	17628
Low	<0	PolII $\sim$ CEBPβ + DNase I HS + H3K4Me3 + H3K4Me1	0.4306	21524
High	>2	PolII $\sim$ CEBPα + H3K4Me3 + DNase I HS + Pu1	0.3651	1714
Med	<2 & >0	PolII $\sim$ H3K4Me3 + DNase I HS + Pu1	0.3737	11400
High	>1.75	PolII $\sim$ H3K4Me3 + DNase I HS + Pu1	0.4154	2727
Med	<1.75 & >0	PolII $\sim$ H3K4Me3 + DNase I HS + Pu1	0.3245	8823
All Data Model	None	PolII $\sim$ H3K4Me3 + H3K4Me1 + DNase I HS	0.7832	50564

Table 2.7: Models for the stratified data using various groupings of genes.

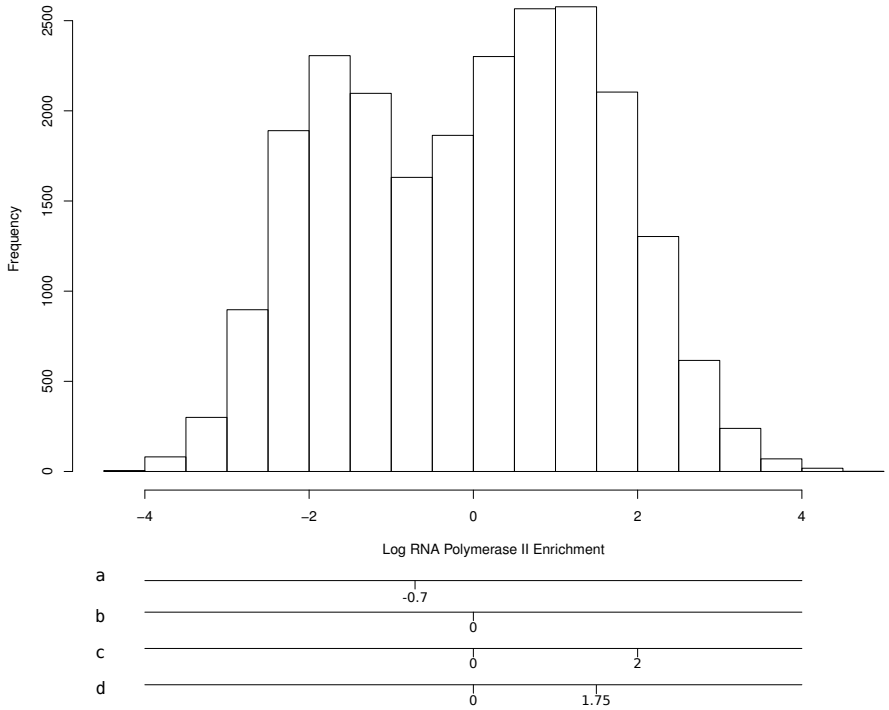


Figure 2.9: Distribution of the log enrichment values for the RNA polymerase II dataset. *a-d* indicate the ranges used for training models on subsets of this data. The cut-off at *a*, -0.7, was based on the lowest point of the bimodal distribution. The cut-off for *b* is based anything below 0 being less than the background noise. *c* and *d* split the region above 0 into two. For *c*, the cut-off is at 2, which arbitrarily assigned as approximately halfway between 0 and the maximum value. For *d*, the cut-off is at 1.75, which splits the number of genes above 0 into two equal groups.

Comparing the cut-offs for genes with high or low expression shows that it is better taking a cut-off based on the two distributions rather than at 0. This is likely due to a statistical, rather than a biological, reason as using a distribution-based cut-off means that more of the signal from the distribution is included in the modeling and produces a better model. Irrespective of this, using 0 or -0.7 as the cut-off for a gene having high or low expression both produce a similar model. The genes classed as having low expression are predicted poorly, with an R^2 of 0.30 and 0.43 using -0.7 and 0, respectively. The genes classed as having high expression are better predicted by the model , with R^2 values of 0.63 and 0.60 using -0.7 and 0, respectively. These higher expressing genes are able to be predicted better by the model because the majority of the information in this region is likely to be actual signal, rather than noise. This means that the model is fitting to, and predicting, the binding of the RNA polymerase II rather than to background noise. Any gene with a log enrichment below 0 has an enrichment value that is below the background noise selected for that region. This means that the modeling for the genes with low expression is being attempted with data with is dominated by background noise, explaining why the models perform poorly. It is interesting to note that when the region above 0 is split into two, regardless of the two cut-offs used, that the models produced perform similarly, but not as well as taking the region as a whole. This is most likely due to the models being produced using “blunt” predictors. All of the models for genes with a log enrichment above 0 result in the same model; RNA polymerase II $\sim$ H3K4Me3 + DNase I Hypersensitivity + Pu1. This model is very interesting as it comprises a histone modification that is known to be strongly indicative of active genes, the DNase I hypersensitivity data which indicates whether the genes are open to be bound by RNA polymerase II, and a transcription factor that plays a large role in the development and maintenance of macrophage function. All three of these predictors are likely to have a fairly coarse-grained effect on gene transcription. None of them are involved in the fine-grained regulation of expression, possibly explaining why the model performs better over a larger range rather than arbitrarily splitting the expressed genes into small sets.

2.4 Discussion

Overall, this work shows that using a simple measure based on read-counts from ChIP-seq data, combined with a simple linear model, it is possible to predict the level of RNA polymerase II binding in mouse macrophage cells. The best model produced using this data was one only containing DNase I hypersensitivity and the histone modifications, H3K4Me3 and H3K4Me1. The lack of the presence of transcription factors indicates that there is a redundancy, at least in a statistical sense, between transcription factors and histone modifications/accessibility when it comes to predicting RNA polymerase II binding. It is interesting to note that the addition of three transcription factors, CEBPα, p65 and PuI had little improvement on the quality of the model and the amount of variance in the RNA polymerase II binding data. Further analysis of models containing transcription factors shows that their addition and the improvement upon the simpler model has little to do with any biological relevance and is based on the model having more datasets to optimise with. The comparison of models with and without transcriptions factors show that the presence of the transcriptions factors generally affects genes that have similarly predicted transcription factor binding sites as a random selection of genes. Analysis of the subsets of the RNA polymerase II enrichment data shows it is far easier to predict the level of RNA polymerase II binding in a broad sense, using these histone modifications/accessibility, but is less accurate when a smaller range of data is used. This is due to the nature of the markers used and the limited amount of data available for the modeling. The markers used for all of the models of the data with log enrichments above 0 were DNase I hypersensitivity, H3K4Me3 and PuI. These are fairly broad indicators of transcription; therefore the wider the range of data used for the modelling, the better the model is. While the transcription factors are important for the transcription of genes, Pu1 especially being an overall regulator of macrophage differentiation and functional maintenance, they are *general* regulators of transcription and the models miss the fine-grained control of the markers themselves and their fine-grained control of transcription. The macrophage-related genes have their enrichment values predicted fairly well by the models produce here. While ideally they would be predicted well enough to be identified at some of the best-predicted genes, the model predicts all of them well enough that none of them are present in the outlying genes. The models don't predict the macrophage-related genes better than any other set of genes as the models were trained using all transcription factors. This means that the models are good at predicting transcription factors in general in macrophage cells, rather than a specific sub-set of transcription factors. This could be addressed by training the models on a sub-set of the genes, such as

the macrophage-related genes or immune-response related genes, then tested on the rest of the data. Doing this could potentially highlight sets of genes which have the same regulatory mechanisms as the set of genes the model was trained on. In practice, this is not as simplistic as it appears as there would need to be a large enough set of genes for the models to be trained on and even having similar functions within the cell does not mean that the regulation of their expression is managed using similar mechanisms. The investigation into the genes that are predicted well or poorly (and thus are likely to have similar or dissimilar regulatory mechanisms, respectively) is likely to be non-trivial due to the large numbers of genes. Recent work has shown that there is a large amount of non-specific transcription factor binding that does not affect transcription.[118] While this was found in yeast, it is likely to hold true for for higher eukaryotes and would go some way to explaining why the transcription factor data added very little to the predictive power of the models over just using the histone modifications. If this non-specific transcription factor binding does not influence gene expression, then the information about transcription factor binding that *does* have an effect on transcription is likely to be masked behind the noise of the binding that doesn't affect transcription. Given the large number of transcription factor binding sites that ChIP-seq experiments find compared to the number of transcribed elements in the genome, the likelihood that there are detected binding sites that do not affect transcription is high.

While it is hard to compare the models constructed here to the models previously developed by Ouyang *et al*[74] and Cheng & Gerstein[105] due to different aims of the models, a comparison can be made. Ouyang *et al* used 12 transcription factors and principle component analysis and were able to explain 70% of the variance in their response variable; expression data from micro-array experiments. Cheng and Gerstein used 12 transcription factors and 7 histone modifications with support vector regression to explain 72% of the variance in their response variable. The model developed here, using the simple enrichment values with linear regression produced a final model of just 3 histone modifications, starting from 5 transcription factors and 3 histone modifications, was able to explain 78% of the variance in the RNA polymerase II binding data. While the model constructed here can explain more of the variance in the expression-metric used, and the comparison between the methods is interesting, it would be unfair to say it performs better than the previous studies due to the different aims of the methods. Ouyang *et al* were specifically using principle component analysis to find cooperation between transcription factors, Cheng and Gerstein using the support vector regression to look at differences in the special patterning of transcription factors and histone modifications - with note of the statistical redundancy between them - whist here the aim was to look

more specifically at a mechanistic hypothesis for RNA polymerase II binding. The redundancy between transcription factors and histone modifications, while not as clear as that found by Cheng and Gerstein, is still evident in the results here. A significant difference between the results here and Cheng's is the histone modifications being dominant in the optimal models. While the method developed here marginally out-performs the methods of Ouyang *et al* and Cheng and Gerstein, it has makes assumptions each of the other methods try to address. Ouyang *et al*, with their transcription factor association score, attempt to take into account more distal transcription factor binding sites that may still have an influence on the expression of a gene. Cheng and Gerstein, by using machine learning, do not make the assumption that the relationship between the transcription factors and expression is a linear one; a method that could potentially improve the models produced here. The method used here does, however, encompass the different binding profiles of the histone modifications that is missed in method of Cheng and Gerstein, and uses histone modification data that was not available to Ouyang *et al*. This is likely one of the sources of improvement over their methods as it means that more of the subtle interaction between the histone modifications and expression is used by the models.

Using the simple metrics in this work has led to a good model for RNA polymerase II binding. It is not, however, comprehensive. Due to only a small number of transcription factors and histone modifications being used, a lot of the fine-grain detail has been lost and only a broad, coarse-grained model is possible. The model produced uses factors that are known to be strong indicators of gene transcription. The over-all mechanism of transcription factor binding is simplified in the resulting model. With this small amount of data is is difficult to investigate the more subtle interplay between the large variety of transcription factors that control the regulation of RNA polymerase II binding.

Chapter 3

Genetic Regulation in Human Stem Cells

3.1 Introduction

3.1.1 Importance of embryonic stem cells

Embryonic stem cells originate in the center of a fertilised egg prior to plantation and are pluripotent cells with the potential to differentiate into any other cell type. They are obtained from the central mass of an embryo before implantation, which normally occurs 4-5 days after fertilisation of the egg. Embryonic stem cells not only have the ability to differentiate into any other cell type but also have a limitless ability to self-renew.[119] The pluripotency of embryonic stem cells is unmatched in any other cell type, and it is this which makes them such an interesting research target. The potential of the stem cells to develop into any other cell type, and react quickly to any differentiation signals they receive, means that tight control must be exerted over gene expression.

Whilst there are stem cells present in adults, they don't have the total potential of embryonic stem cells; for example bone marrow derived stem cells are unable to form cells of the central nervous system. Embryonic stem cells are an important research target for pluripotency and lineage commitment, but their source has caused some controversy. There is work being done to convert adult stem cells into embryonic stem cell-equivalent cells, though there is a vast amount of work to be done in this field.[120] Research is being done into using stem cells to treat a number of pathologies including diabetes,[121] Parkinson's,[122;123] cancer,[124] cardiovascular disease,[125] and Alzheimer's disease.[126]

The capacity of embryonic stem cells to differentiate into any cell type, while retaining full capacity for self-renewal, makes them an incredibly interesting research subject. We are only beginning to understand their potential, how they function, and how they maintain their pluripotency.

3.1.2 Transcriptional control of stem cells

There are three core transcription factors that exert the majority of the control over the pluripotency and self-renewing characteristics of human and mouse embryonic stem cells: Oct4/Pou5F1, Sox2 and Nanog.[127]

Oct4 is essential for pluripotency and is only expressed in pluripotent and totipotent cells.[128;129] The level of Oct4 expression must be highly controlled as either over- or under-expression leads to differentiation.[52] Sox2 is also essential for pluripotency.[130] Sox2 and Oct4 both regulate embryonic stem cell-specific genes, including themselves.[131–139] The third of the core

embryonic stem cell proteins, Nanog, is only required to form pluripotent cells.[140] Once the pluripotency is reached, cells in which Nanog has been silenced maintain pluripotency but are more likely to spontaneously differentiate.[141;142] Further evidence of the increased likelihood of differentiation of Nanog-deficient cells is under conditions causing increased DNA damage. In this situation, p53 suppresses expression of Nanog and results in a more differentiated state, which in turn allows for p53-dependent cell cycle arrest and apoptosis.[143;144]

The core embryonic stem cell transcription factors are often found co-binding at active and silent locations across the genome.[145;146] There have been studies into the co-binding sites of these core transcription factors which have highlighted various interesting targets, including other genes required for pluripotency and the encoding of proteins required for the modification of chromatin.[147–154] The manner in which they repress and promote expression at each of these locations is still poorly understood.

It has also been found that there is a second important set of transcription factors which influence embryonic stem cells - the Myc cluster.[155] The Myc cluster, which includes c-Myc, n-Myc, Rex1, Zfx and E2f1, is thought to be independent of Oct4 regulation. It is known that this cluster is involved in metabolism, as well as self-renewal.[156–158] The extent of the binding in this cluster is vast, with almost a third of all active embryonic stem cell genes bound by the core transcription factors and c-Myc.[159] It is thought that the core transcription factors work to recruit the transcriptional machinery to the genes, while c-Myc is responsible for the control of the pause-release of this machinery.[159] Recent work, however, has postulated that the function of the Myc cluster acts by amplifying the effect of genes that are already transcribed, rather than activating new genes.[160;161]

Overall, the maintenance of pluripotency and limitless self-renewal is a fine balance between a large number of transcription factors and the action and expression of each being under tight control by the master regulators, Oct4, Nanog and Sox2. While the exact function and interaction of all the transcription factors in the cell is still unknown, much work is being done in this field.

3.1.3 Histone modification control of stem cells

In embryonic stem cells, histone modifications are as tightly controlled as transcription factor binding. Having the correct histone modifications in the appropriate places is essential for the maintenance of pluripotency and self-renewal.

Generally, H3K4Me3 presence is directly related to RNA polymerase II binding and tran-

scription.[162] In embryonic stem cells this is not the case; most of the promoters across the genome are marked with H3K4Me3, whether the gene they are associated with is transcribed or not.[163;164] The balance of methylation marks at lysine 4 of histone H3 is a delicate affair in stem cells: while H3K4Me3 is present at most promoters it is often found alongside the repressive H3K27Me3 modification which results in the gene being in a poised state.[54] This poised state very often occurs for genes that are essential for differentiation.[54;55;165–167] The presence of the H3K27Me3 marker silences the locations bound by H3K4Me3 while the cell is still pluripotent, but upon differentiation, one of the two markers is removed, leaving either an active or silenced gene depending on the fate of the cell.[54;55;165] The purpose of this combination of activation and repression markers is thought to be to allow rapid response to differentiation signals that the stem cell receives. The role of H3K27Me3 is complicated and still not entirely clear. H3K27Me3 has been implicated in the poised state signalling,[54;55;165] in regulating pluripotency,[168] and even cellular reprogramming,[169] though its full role is still being studied. The balance of H3K4Me mono-, di- and tri- methylation is essential for the careful maintenance of stem cell physiology. Cells with a deficiency in H3K4Me1/2 specific de-methylases spontaneously differentiate[170] and cells lacking H3K4Me2/3 specific de-methylases lose the ability to self-renew.[171]

H3K9Me3, a long-range silencing marker, plays an important role in the regulation of stem cell pluripotency. H3K9Me3 regions have been found to be much smaller, and the effect of its long-range silencing increased,[172] in stem cells as compared to differentiated cells.[173] Despite increased effects in differentiated cells, H3K9Me3 has been shown to regulate Oct4 down-regulation by promoting DNA methylation of the Oct4 gene.[174] The histone de-methylases Kdm3a and Kdm4c regulate the H3K9Me3 marker: the loss of these proteins results in non-reversible stem cell differentiation, highlighting the importance of this marker.[175]

H3K79 methylation is normally found within the gene body of actively transcribed genes.[176] Depletion of Kmt4, the methyltransferase responsible for H3K79 methylation, in stem cells results in a lower proliferation rate while maintaining pluripotency and self-renewal abilities.[177] Interestingly, depletion of Kmt4 also results in more open chromatin due to the loss of two histone modifications, H3K9Me2 and H4K20Me3, at the centromeres and telomeres[177] as well as increased expression of Nanog, though the mechanism for this is unclear.[178]

Histone acetylation opens up chromatin by neutralising the positive charge of the lysine residues. Histone acetylation is very common in embryonic stem cells compared with differentiated cells, highlighting the open nature of the chromatin in these embryonic stem cells.[162;179–181] Inhibiting histone de-acetylases in stem cell results in cells with a reduced ability to differentiate

and stronger self-renewal.[182;183]

3.1.4 ENCODE

The Encyclopedia of DNA Elements (ENCODE) builds on the work of the human genome project. The human genome project sequenced the whole of the human genome for the first time. While the objective of the human genome project was to sequence the DNA in a single person, the objective of ENCODE is to find and characterise the functional elements of the sequence.[184] The human genome project paved the way for ENCODE, establishing a reference sequence. The established base is that $\sim$23000 protein-coding genes of the human genome make up only 1.5% of the $\sim$3 billion base pairs. The aim of ENCODE is to establish the function of the rest of our DNA, which includes many regulatory elements, non-protein coding elements and supposed "junk", non-functional, DNA. The first wave of publications from the ENCODE project were released on the 5th of September 2012; at that point, there were 1640 datasets available from 147 different cell lines.[185] The initial findings debunked the idea of "junk" DNA, showing that 80% of the genome can be classified as having biological function (protein/non-coding RNA or a reproducible biochemical signature). In at least one cell type, 95% of the genome is within 8kb of a protein-DNA interaction and 99% of it is within 1.7kb of a “biochemical event”.[185]

Overall, the ENCODE project and the accessibility of the data it produced are a massive step forwards in the understanding of the function of the human genome and a boon for the scientific community.

3.1.5 Aims and Objectives

The aim of this chapter is to take the methods previously used on the mouse macrophage data and apply them to a larger set of data derived from embryonic stem cells. Using the same simple enrichment value, LASSO regression was used to construct models using 24 histone modification and 23 transcription factor datasets. As previously, RNA polymerase II binding was taken as an equivalent to gene expression levels; the correlation between the two has been previously shown.[74;75]

3.2 Methods

3.2.1 Data sets

The datasets used were obtained from the GEO database.[106] All of the datasets used were from ChIP-seq experiments. Data from the human embryonic cell line H1 (table 3.1) were compared to data from the H9 (table 3.2) and IMR90 (table 3.3) stem cell lines. As per the ENCODE requirements, all of these cell lines were grown in the same conditions and prepared and sequenced in a similar manner. If the datasets were comparable, where a marker appeared in multiple cell lines, the markers would be highly correlated in all cases and clustering would occur between identical markers of different cell lines, rather than within cell lines. For the comparison of the datasets all available data were used, but for subsequent modelling only used one dataset for each marker: the dataset with the highest number of reads. The datasets were imaged and clustered using the default arguments for the *heatmap* function of the R Statistical package.[112] Models were constructed for all four of the RNA Polymerase II datasets available for the H1 cell line. It was decided that modeling each RNA polymerase II dataset was better than selecting one or taking some form of mean or consensus of the four sets together. Analysing four models allows for a comparison of the models, giving a consensus as to what predictors are eliminated and better overview as to the mechanistic hypothesis for RNA Polymerase II binding. As there are multiple datasets for the same marker, only the dataset with the most reads was used in the modelling stages. These datasets are indicated by the entries in bold in table 3.1.

Accession	Marker	Marker Type
SRR351841	ATF3	Transcription Factor
SRR351842	ATF3	Transcription Factor
SRR351776	BALIIA	Transcription Factor
SRR351628	BCLIIA	Transcription Factor
SRR351670	CTCF	Transcription Factor
SRR351671	CTCF	Transcription Factor
SRR351691	EGR-1	Transcription Factor
SRR351692	EGR-1	Transcription Factor
SRR351600	FOSL1	Transcription Factor
SRR351601	FOSL1	Transcription Factor

Continued on next page.

Accession	Marker	Marker Type
SRR351679	GABP	Transcription Factor
SRR351680	GABP	Transcription Factor
GSM602257	H2AK5ac	Histone Modification
GSM602258	H2AK5ac	Histone Modification
GSM605295	H2BK120ac	Histone Modification
GSM605296	H2BK12ac	Histone Modification
GSM605297	H2BK12ac	Histone Modification
GSM605298	H2BK15ac	Histone Modification
GSM605299	H2BK15ac	Histone Modification
GSM605300	H2BK20ac	Histone Modification
GSM605301	H2BK20ac	Histone Modification
GSM605302	H2BK5ac	Histone Modification
GSM605303	H2BK5ac	Histone Modification
GSM602259	H3K18ac	Histone Modification
GSM605304	H3K18ac	Histone Modification
GSM667614	H3K18ac	Histone Modification
GSM667615	H3K18ac	Histone Modification
GSM667617	H3K23ac	Histone Modification
GSM667618	H3K23ac	Histone Modification
GSM605305	H3K23me2	Histone Modification
GSM605306	H3K23me2	Histone Modification
GSM466732	H3K27ac	Histone Modification
GSM663427	H3K27ac	Histone Modification
GSM434776	H3K27me3	Histone Modification
GSM466734	H3K27me3	Histone Modification
GSM605308	H3K27me3	Histone Modification
GSM409312	H3K36me3	Histone Modification
GSM450268	H3K36me3	Histone Modification
GSM466737	H3K36me3	Histone Modification
GSM605309	H3K36me3	Histone Modification

Continued on next page.

Accession	Marker	Marker Type
GSM605311	H3K4ac	Histone Modification
GSM667624	H3K4ac	Histone Modification
GSM409307	H3K4me1	Histone Modification
GSM434762	H3K4me1	Histone Modification
GSM466739	H3K4me1	Histone Modification
GSM605312	H3K4me1	Histone Modification
GSM602260	H3K4me2	Histone Modification
GSM602261	H3K4me2	Histone Modification
GSM409308	H3K4me3	Histone Modification
GSM469971	H3K4me3	Histone Modification
GSM605315	H3K4me3	Histone Modification
GSM605317	H3K56ac	Histone Modification
GSM667627	H3K56ac	Histone Modification
GSM605318	H3K79me1	Histone Modification
GSM605319	H3K79me1	Histone Modification
GSM605320	H3K79me1	Histone Modification
GSM605321	H3K79me2	Histone Modification
GSM605322	H3K79me2	Histone Modification
GSM434785	H3K9ac	Histone Modification
GSM605323	H3K9ac	Histone Modification
GSM428291	H3K9me3	Histone Modification
GSM605325	H3K9me3	Histone Modification
GSM605327	H3K9me3	Histone Modification
GSM605328	H3K9me3	Histone Modification
GSM605329	H4K20me1	Histone Modification
GSM605330	H4K5ac	Histone Modification
GSM605332	H4K91ac	Histone Modification
SRR351523	HDAC2	Transcription Factor
SRR351524	HDAC2	Transcription Factor
SRR351876	JunD	Transcription Factor

Continued on next page.

Accession	Marker	Marker Type
SRR351877	JunD	Transcription Factor
SRR351702	NANOG	Transcription Factor
SRR351703	NANOG	Transcription Factor
SRR351566	NRSF	Transcription Factor
SRR351567	NRSF	Transcription Factor
SRR351902	p300	Transcription Factor
SRR351903	p300	Transcription Factor
SRR351568	Pol II	RNA Polymerase II
SRR351569	Pol II	RNA Polymerase II
SRR351787	Pol II	RNA Polymerase II
SRR351788	Pol II	RNA Polymerase II
SRR351704	POU5F1	Transcription Factor
SRR351705	POU5F1	Transcription Factor
SRR351758	RAD21	Transcription Factor
SRR351759	RAD21	Transcription Factor
SRR351565	Rev X link Chromatin	Transcription Factor
SRR351829	RXRA	Transcription Factor
SRR351830	RXRA	Transcription Factor
SRR351687	Sin3AK-20	Transcription Factor
SRR351688	Sin3AK-20	Transcription Factor
SRR351645	SIX5	Transcription Factor
SRR351646	SIX5	Transcription Factor
SRR351590	SP1	Transcription Factor
SRR351591	SP1	Transcription Factor
SRR351681	SRF	Transcription Factor
SRR351682	SRF	Transcription Factor
SRR351727	TAF1	Transcription Factor
SRR351728	TAF1	Transcription Factor
SRR351819	TAF7	Transcription Factor
SRR351820	TAF7	Transcription Factor

Continued on next page.

Accession	Marker	Marker Type
SRR351685	TCF12	Transcription Factor
SRR351686	TCF12	Transcription Factor
SRR351683	USF-1	Transcription Factor
SRR351684	USF-1	Transcription Factor
SRR351843	YY1-(sc-281)	Transcription Factor
SRR351844	YY1-(sc-281)	Transcription Factor

Table 3.1: The datasets used for the generation of models for human H1 embryonic stem cell line from the GEO database.[106] Entries marked in bold were later used for the linear models.

Accession	Marker	Marker Type
GSM605307	H3K27ac	Histone Modification
GSM605310	H3K36me3	Histone Modification
GSM605314	H3K4me2	Histone Modification
GSM605316	H3K4me3	Histone Modification
GSM605324	H3K9ac	Histone Modification
GSM616127	H3K4me2	Histone Modification
GSM616128	H3K4me3	Histone Modification
GSM616129	H3K9ac	Histone Modification
GSM665037	H3K27ac	Histone Modification
GSM667608	H2AK5ac	Histone Modification
GSM667609	H2AK5ac	Histone Modification
GSM667610	H2BK12ac	Histone Modification
GSM667611	H2BK12ac	Histone Modification
GSM667612	H2BK5ac	Histone Modification
GSM667613	H2BK5ac	Histone Modification
GSM667616	H3K18ac	Histone Modification
GSM667619	H3K23ac	Histone Modification
GSM667620	H3K23ac	Histone Modification
GSM667621	H3K23me2	Histone Modification
GSM667622	H3K27me3	Histone Modification
GSM667623	H3K36me3	Histone Modification
GSM667625	H3K4ac	Histone Modification
GSM667626	H3K4me1	Histone Modification
GSM667628	H3K56ac	Histone Modification
GSM667629	H3K79me1	Histone Modification
GSM667630	H3K79me2	Histone Modification
GSM667631	H3K9me3	Histone Modification
GSM667632	H3K9me3	Histone Modification
GSM667633	H3K9me3	Histone Modification
GSM667634	H4K20me1	Histone Modification
GSM667635	H4K5ac	Histone Modification
GSM667636	H4K5ac	Histone Modification
GSM667637	H4K8ac	Histone Modification
GSM667638	H4K8ac	Histone Modification
GSM667639	H4K91ac	Histone Modification
GSM667640	H4K91ac	Histone Modification
GSM667643	Input	Histone Modification

Table 3.2: The datasets used for the generation of models for human H9 embryonic stem cell line from the GEO database.[106]

Accession	Marker	Marker Type
GSM605307	H3K27ac	Histone Modification
GSM605310	H3K36me3	Histone Modification
GSM605314	H3K4me2	Histone Modification
GSM605316	H3K4me3	Histone Modification
GSM605324	H3K9ac	Histone Modification
GSM616127	H3K4me2	Histone Modification
GSM616128	H3K4me3	Histone Modification
GSM616129	H3K9ac	Histone Modification
GSM665037	H3K27ac	Histone Modification
GSM667608	H2AK5ac	Histone Modification
GSM667609	H2AK5ac	Histone Modification
GSM667610	H2BK12ac	Histone Modification
GSM667611	H2BK12ac	Histone Modification
GSM667612	H2BK5ac	Histone Modification
GSM667613	H2BK5ac	Histone Modification
GSM667616	H3K18ac	Histone Modification
GSM667619	H3K23ac	Histone Modification
GSM667620	H3K23ac	Histone Modification
GSM667621	H3K23me2	Histone Modification
GSM667622	H3K27me3	Histone Modification
GSM667623	H3K36me3	Histone Modification
GSM667625	H3K4ac	Histone Modification
GSM667626	H3K4me1	Histone Modification
GSM667628	H3K56ac	Histone Modification
GSM667629	H3K79me1	Histone Modification
GSM667630	H3K79me2	Histone Modification
GSM667631	H3K9me3	Histone Modification
GSM667632	H3K9me3	Histone Modification
GSM667633	H3K9me3	Histone Modification
GSM667634	H4K20me1	Histone Modification
GSM667635	H4K5ac	Histone Modification
GSM667636	H4K5ac	Histone Modification
GSM667637	H4K8ac	Histone Modification
GSM667638	H4K8ac	Histone Modification
GSM667639	H4K91ac	Histone Modification
GSM667640	H4K91ac	Histone Modification
GSM667643	Input	Histone Modification

Table 3.3: The datasets used for the generation of models for human IMR90 embryonic stem cell line from the GEO database.[106]

3.2.2 Enrichment Calculation

Enrichment values were calculated for each gene within each dataset using the same method as described in section 2.2.2. The background values used in the calculation of the enrichment value are as described in section 2.2.2. The enrichment values used are the number of reads for the gene and flanking region (2000bp up and down-stream of the transcription start and end sites) divided by a background value of reads per base multiplied by the length of the gene and flanking region. For the background value we used the highest of the mean reads per base for the whole of the dataset, the chromosome the gene was on or two 10,000bp regions surrounding the gene/flanking region. As previously, for genes where the enrichment value was 0, the value was set to 0.001 so that it was possible to take the log for the gene. In this case only the gene and flanking regions together were considered for the enrichment values. This was in part owing to the results of the previous chapter finding where using regions together led to more robust results, and partly owing to the prevalence of the "paused" histone state. This "paused" state occurs often in embryonic stem cells and is marked by the occurrence of both the H3K4Me3 and H3K27Me3 modifications.[54;163;164]

3.2.3 Linear Models

Linear models were employed to fit and optimise the transcription factor and histone modification data to the RNA polymerase II binding data. Using linear models it is possible to remove predictors from the model and re-run them. It is then possible to compare different models to identify which of them are better at fitting to the RNA polymerase II binding data, and thus are more likely to represent reality. Owing to the large number of datasets available, the simple forward stepwise elimination of predictors is not a viable method for optimising a linear model, due to the large number of possible combinations of predictors. For the model selection, LASSO (least absolute shrinkage and selection operator) regression was used.[186] LASSO regression is a regularised least-squares regression method that calculates a standard regression model (equation 3.1) with a restriction that the parameter vector (L^1-norm) is less than a tuning parameter (s, equation 3.2). LASSO regression is a method for selecting a sub-set of predictors that result in the best fit for the response variable given the cutoff used. Using this form of regression not only allows the removal of redundant datasets from the model but also, in doing this, helps to avoid over-fitting to the response variable (RNA polymerase II

binding in this case).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots \beta_p x_{ip}, \qquad i = 1, \ldots, n, \tag{3.1}$$

$$\min \sum (y - \hat{y})^2 \text{ subject to } \sum |b_p| \leq s \tag{3.2}$$

Where y is the observed data, x is the predictor, b the coefficient and s is the chosen tuning or constraint parameter. As the model is calculated, choosing s, the constraint value, is equivalent to choosing the number of predictors to be used in the final model.

Cross-validation was used to calculate the optimal constraint value for the dataset. Fifty-fold cross-validation was used. The Least Angle Regression (LARS) algorithm[187] was used to perform the LASSO regression for each step of the cross-validation and the final constraint value used was within one standard error of the mean. Using this constraint value allowed elimination of the highest number of predictors, whilst retaining a reasonable degree of error.

3.3 Results

3.3.1 Comparison of Cell Lines

A comparison of all the available datasets for the H1, H9 and IMR90 cell lines was done to assess whether the data from each cell line were comparable. The aim of this was to establish whether the datasets were usable together so as to expand the data available for the later modeling of RNA polymerase II binding. Figure 3.1 shows the correlation of the enrichment values for each gene for all available data for these three cell lines. Here, it is obvious that the datasets from IMR90 cells (green) are mostly clustered together, indicating that the IMR90 datasets correlate better with themselves rather than with identical markers from the other cell lines. The clustering of the IMR90 datasets contradicts combining datasets from multiple cell lines into a single linear model. A closer comparison of the H1 and H9 cell lines was required to establish whether it was feasible to use a combination of the two cell lines in the modelling stages (table 3.2). For the most part it appears that the data would be comparable, but there is a significant amount of clustering of the H9 cell line data in one part of the plot. This cluster together accounts for a third of the H9 datasets (13 of 37). While this cluster of H9 datasets were correlated to some extent with a fair number of the H1 datasets it was decided that using a single cell line would be the best approach in modelling the binding of RNA polymerase II so as to avoid unnecessary noise or biologically irrelevant datasets to the data used in the models. Figure 3.3 shows the correlation between the enrichment values for the histone modifications of the H1 cell line data. The heat map shows the relationship between the modifications that occurring on different histone sub-units and the various modifications which occur. It is evident that the datasets available are enriched in modifications that occur on the H3 protein and that there are a large number of acetylation datasets available. There is a fairly distinct correlation between the majority of the acetylated markers (shown in orange). Figure 3.4 shows the correlations of the enrichment values for the transcription factor datasets for the H1 cell line. This figure shows clear correlations (red) between some transcription factors and also distinct anti-correlations (blue). The four RNA polymerase II datasets are clustered together, as would be expected.

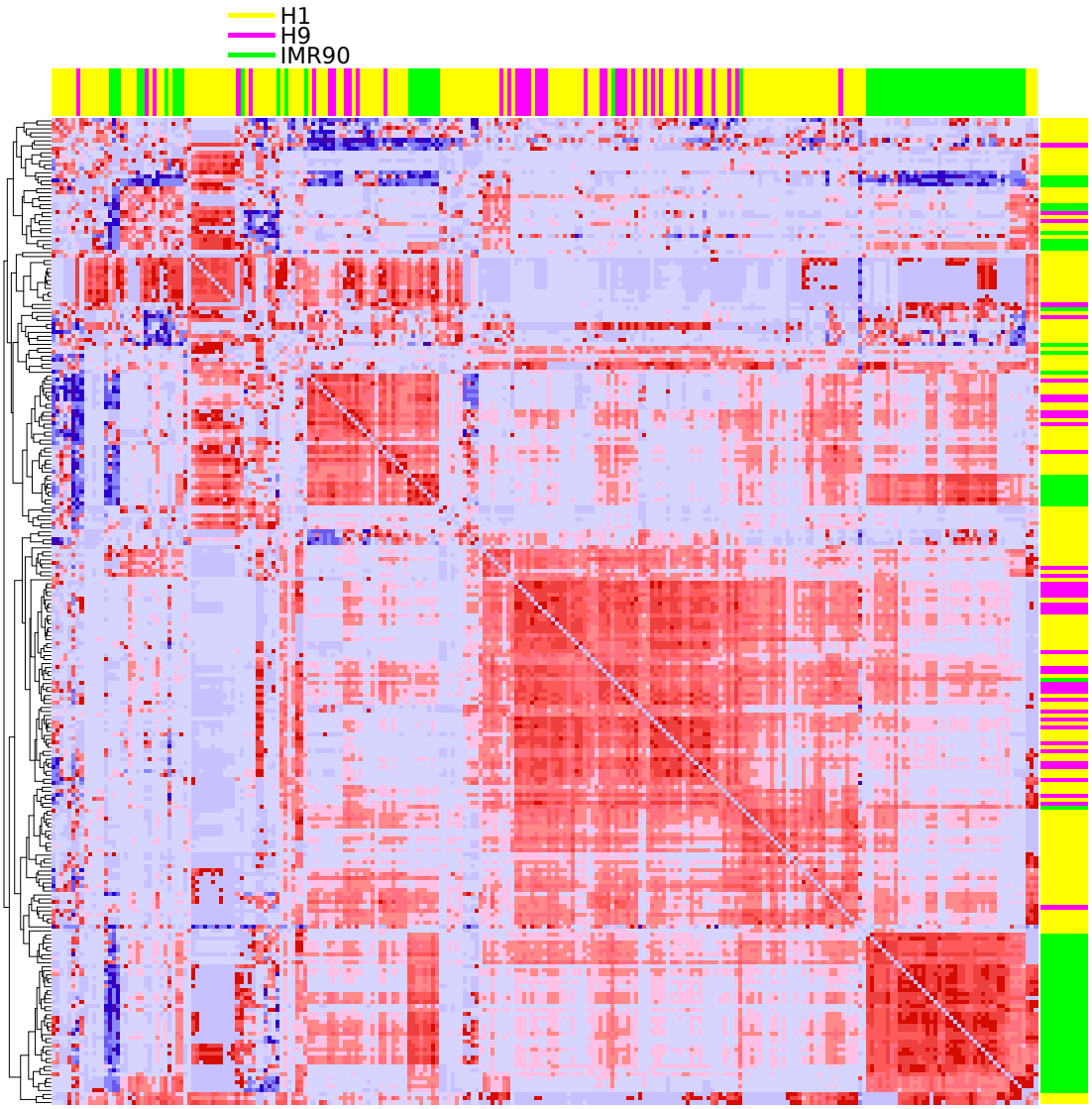


Figure 3.1: Comparison of the correlation between the enrichment scores for three cell lines: H1, H9 and IMR90. Data from H1 cell lines are shown along the yellow lines; H9 cell lines pink lines and IMR90 along the green lines. Squares in dark red are highly correlated, with dark blue negatively correlated.

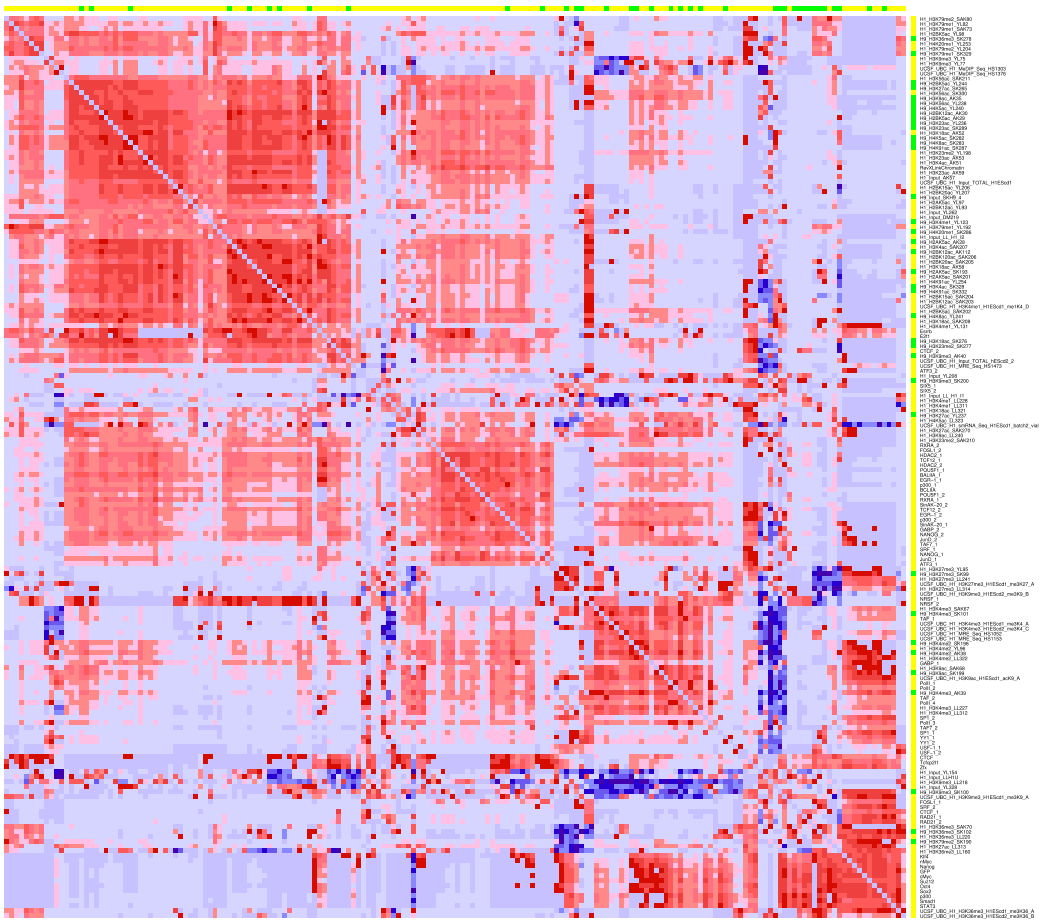


Figure 3.2: Comparison of the correlation between the enrichment scores for the H1 (yellow) and H9 (green) cell lines. Squares in dark red are highly correlated, in dark blue negatively correlated.

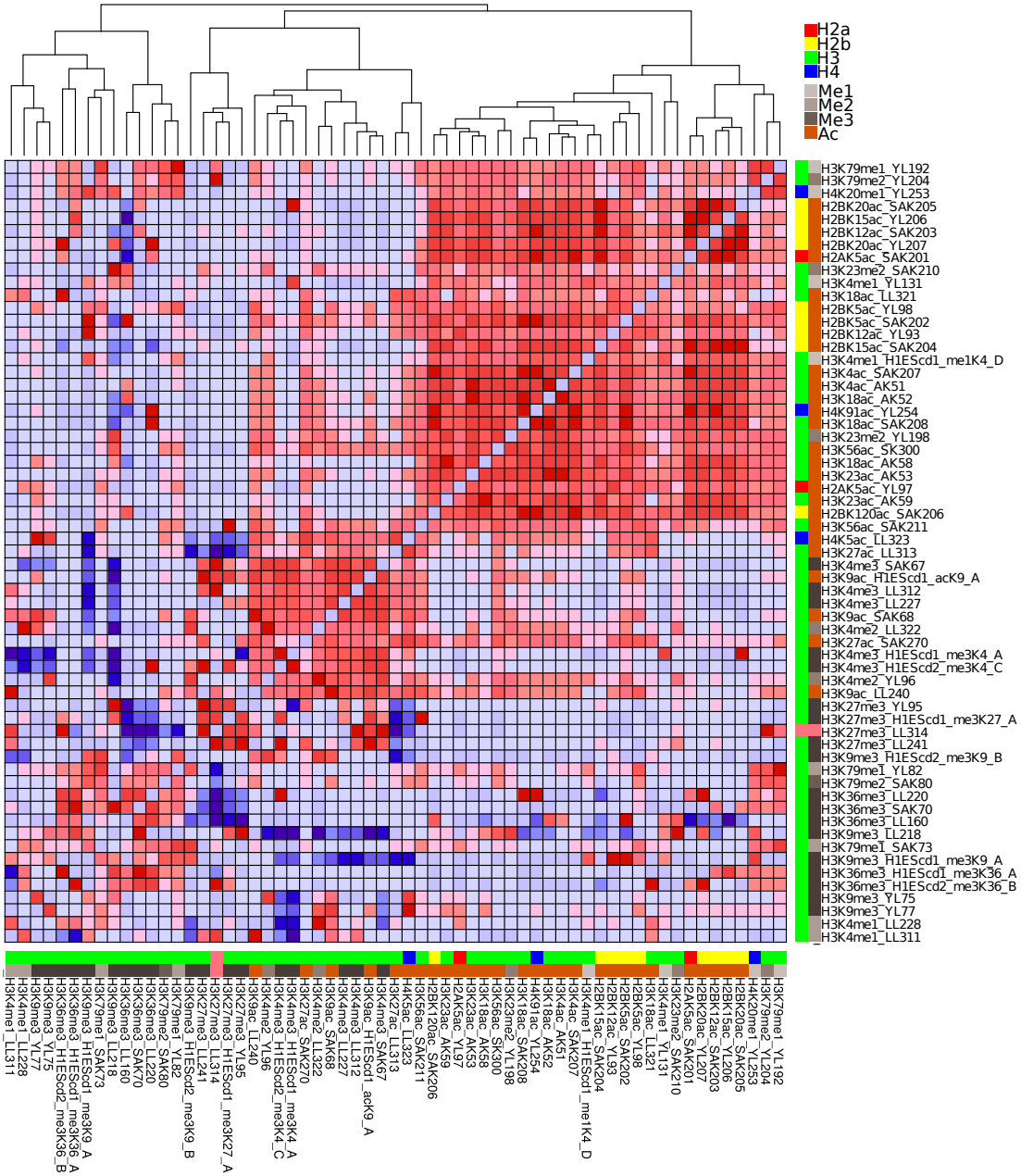


Figure 3.3: Heatmap showing the correlation between the enrichment values for the histone markers for the H1 human embryonic stem cell line. Histone modifications that occur are marked as follows; histone H2a red, H2b yellow, H3 green and H4 blue. Histone modifications that involve mono-methylation (Me1) are marked as light grey, bi-methylation (Me2) as medium grey, tri-methylation (Me3) as dark grey and acetylation (Ac) as orange. Squares in dark red are highly correlated, while those in dark blue are negatively correlated.

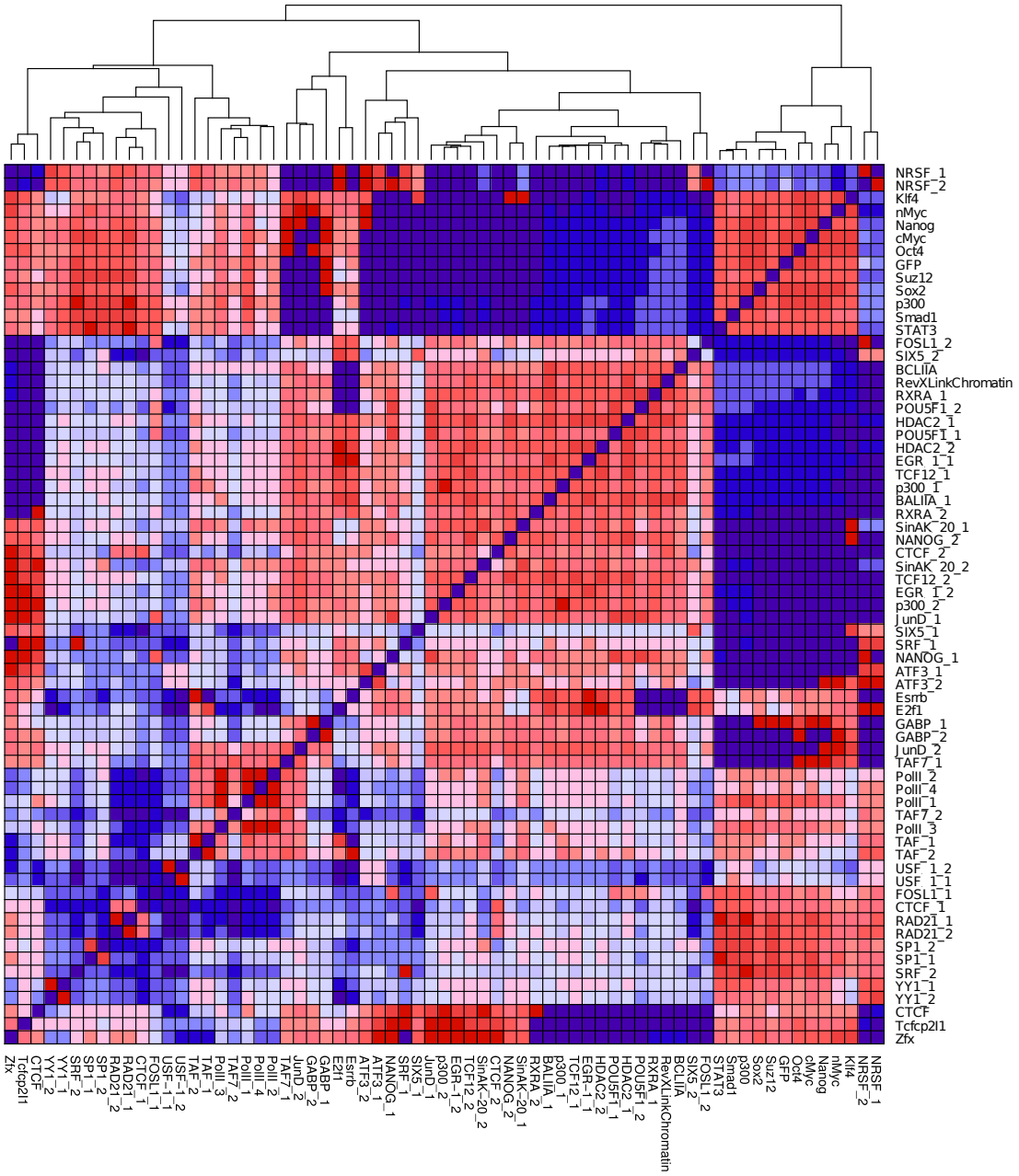


Figure 3.4: Heatmap showing the correlation between the enrichment values for the transcription factors for the H1 human embryonic stem cell line. Squares in dark red are highly correlated, while those in dark blue are negatively correlated.

3.3.2 LASSO regression

Linear regression was performed on the data for the H1 human embryonic stem cell line using the LASSO predictor selection method with the aim of modelling RNA polymerase II binding. Firstly this was done on the enrichment values (figure 3.5, table 3.4), and then on the log of the enrichment values (figure 3.6, table 3.4).

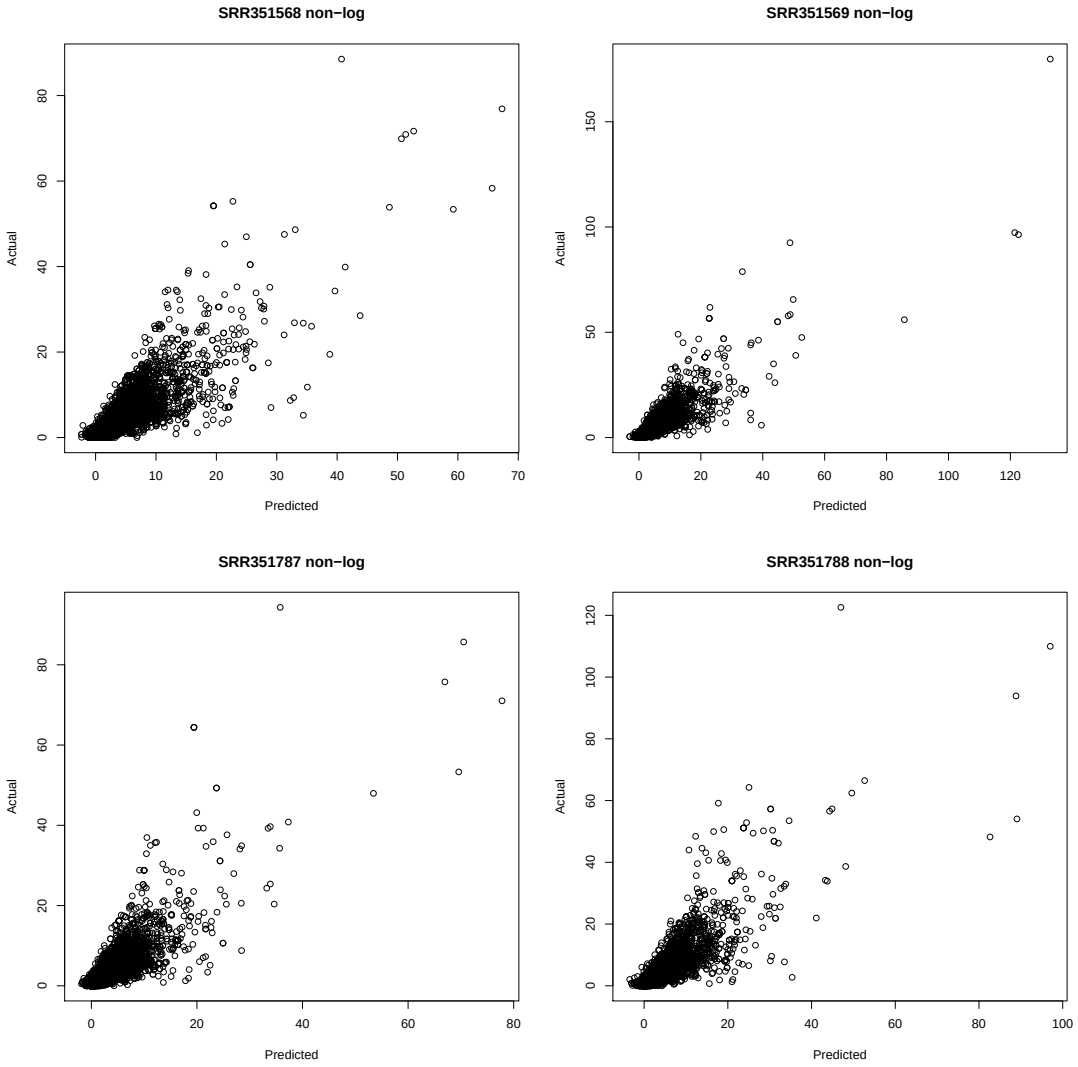


Figure 3.5: The linear models for the four H1 embryonic stem cell RNA Polymerase II datasets. The *actual* data are the enrichment values for each gene for RNA polymerase II. The *predicted* data are the enrichment data for all of the transcription factor and histone modification datasets, scaled by the co-efficients determined by the LASSO regression.

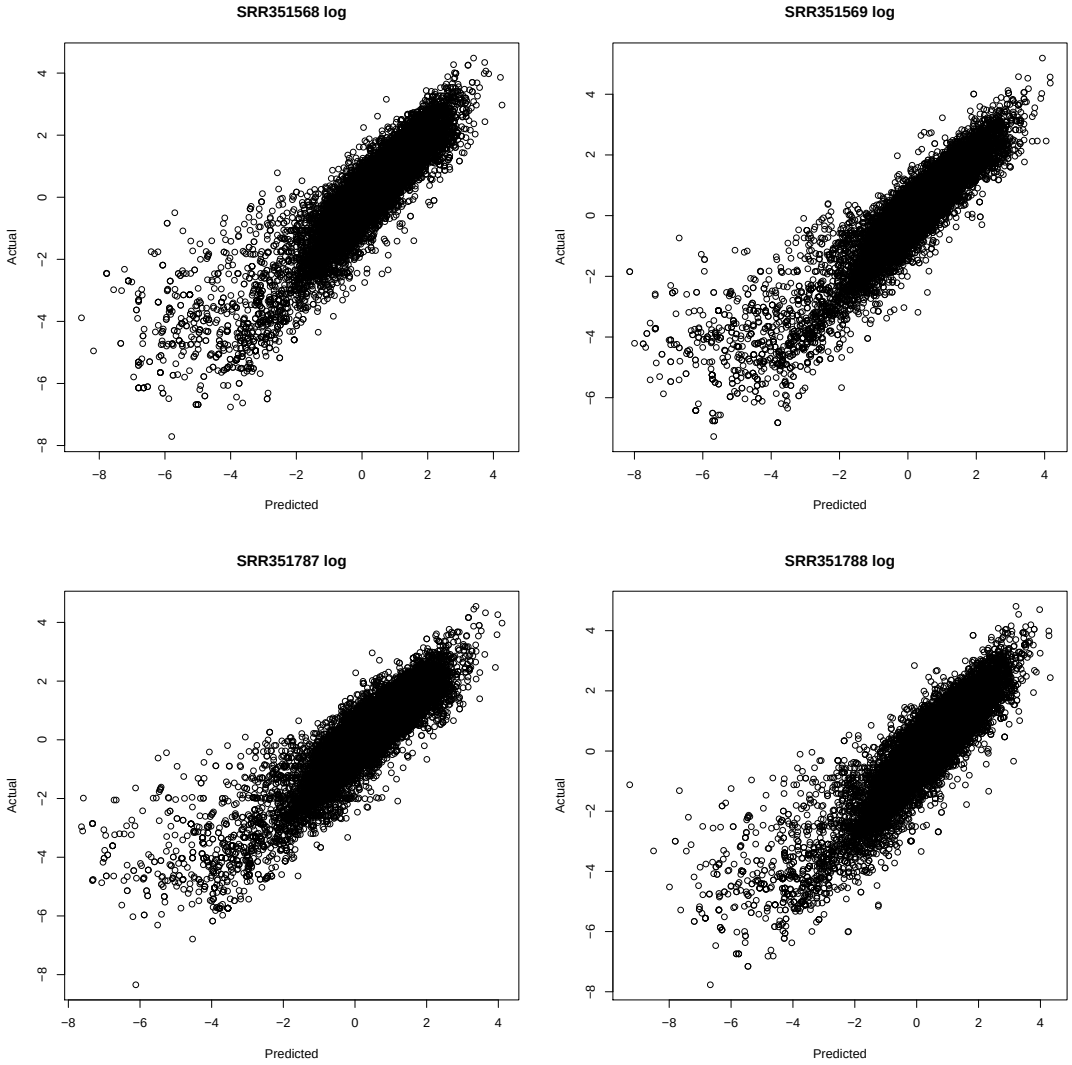


Figure 3.6: The log linear models for the four H1 embryonic stem cell RNA Polymerase II datasets. The *actual* data are the logged enrichment values for each gene for RNA polymerase II. The *predicted* data are the logged enrichment data for all of the transcription factor and histone modification datasets, scaled by the co-efficients determined by the LASSO regression.

Dataset	R^2	Number of predictors used	Number of predictors eliminated	s
SRR351568	0.751	43	4	$1.41x10^{-5}$
SRR351569	0.794	35	12	$1.90x10^{-5}$
SRR351787	0.727	40	7	$2.12x10^{-5}$
SRR351788	0.741	44	3	$7.76x10^{-6}$
SRR351568 log	0.817	40	7	$7.70x10^{-6}$
SRR351569 log	0.843	42	5	$3.61x10^{-6}$
SRR351787 log	0.797	37	10	$1.05x10^{-5}$
SRR351788 log	0.791	46	1	$1.08x10^{-6}$

Table 3.4: Summary of the LASSO models for the four RNA polymerase II datasets for the enrichment values and logged enrichment values. *s* is the selection criterion determined by cross-validation of the LASSO regression models.

The LASSO regression models for both the logged enrichment and non-logged enrichment values were very good. The non-logged models have an R^2 between 0.727 and 0.794 and eliminate between 3-12 predictors. The logged models have an R^2 between 0.791 and 0.843 and eliminate between 1 and 10 predictors. Overall, comparing the logged and non-logged values, the logged values have a higher R^2, though there is a difference between the number of predictors used. It is likely that the increase in R^2 when using the logged enrichment values as using the logs ensures a more linear relationship between the data (table 3.4).

Dataset	R^2	Number of predictors used	Number of predictors eliminated	s
SRR351568 TF	0.720	23	0	0.0
SRR351569 TF	0.757	19	4	$1.1x10^{-5}$
SRR351787 TF	0.693	22	1	$3.15x10^{-6}$
SRR351788 TF	0.686	22	1	$8.19x10^{-6}$
SRR351568 HM	0.683	23	1	$2.89x10^{-6}$
SRR351569 HM	0.697	19	5	$8.71x10^{-6}$
SRR351787 HM	0.647	23	1	$3.70x10^{-6}$
SRR351788 HM	0.639	24	0	0.0

Table 3.5: Summary of the LASSO models for sub-sets of the four RNA polymerase II datasets for the enrichment values and logged enrichment values. *s* is the selection criterion determined by cross-validation of the LASSO regression models. TF: Transcription factors only. HM: Histone Modifications only.

The logged enrichment datasets were then split into two groups: the histone modifications and the transcription factors. Looking at the separate effects of the histone modifications and transcription factors is the simplest way to look at how they affect RNA polymerase II binding and how they act independently of each other. Comparing the R^2 for the transcription factor- or histone modification-only, the transcription-factor models have an R^2 of between 0.686 and 0.757, while the histone modifications have an R^2 of between 0.639 and 0.697. Overall, the transcription factor-only models perform better than the histone modification-only models but

they both eliminate a similar number of predictors in the optimal model. This is interesting and a contrast to the results found in the previous chapter, where the histone modifications alone produced a better model than using transcription factors alone or a mixed transcription factor and histone modification model. For this data, the mixed transcription factor and histone modification model produces a better model than using either of the sub-sets separately. With the exception of one of the RNA polymerase II datasets used, the all-data model has a higher R^2 and uses fewer predictors. The only exception is the SRR351569 RNA polymerase II dataset which eliminates five of the predictors for the histone modification-only sub-set, four of the predictors for the transcription factor-only sub-set but only eliminates four of the datasets in the logged enrichment value model for all of the data. For the non-logged all-data model of the SRR351568 dataset however, 12 of the predictors have a co-efficient of 0 and are eliminated from the model.

Looking more closely at the models and the elimination of predictors, it should be possible to gain some insight for a mechanistic hypothesis of RNA polymerase II binding in human embryonic stem cells. The overall view of the binding of RNA polymerase II is a lot less clear from these models than for the mouse models in the previous chapter.

Taking all of the sub-sets of data, including the non-logged, logged, transcription factor-only and histone modification-only models, there is little overlap between the predictors which are selected out of the models (figure 3.7). For each of these sub-sets of data, there were no occurrences where two sub-sets dropped the same predictor more than twice in each sub-set. Only one predictor, USF-1, was dropped in more than two of the sub-sets, being left out of two non-logged models, one log model and one transcription-factor only model. USF-1 is a transcription factor that has been shown to bind to FOSL1, this elimination could be due to its redundancy with this transcription factor. A comparison of the histone modification-only with the log models show only a single overlap between the predictors that are eliminated: H3K56ac. There is more coordination between the transcription factor-only and the log model however, with three of the six markers that are selected out of the transcription factor-only models also being selected out of the models for the same RNA polymerase II dataset in the log models: GABP, USF-1 and p300. Overall there is little consistency between the sub-sets of the data with few of the markers being selected out multiple times within multiple sub-sets of the data.

Looking closer at each of the sub-sets, there is more consistency within the non-logged and logged models than between them. This does not hold true for the histone modification-only or transcription factor-only sub-sets of data though, here there are no markers selected

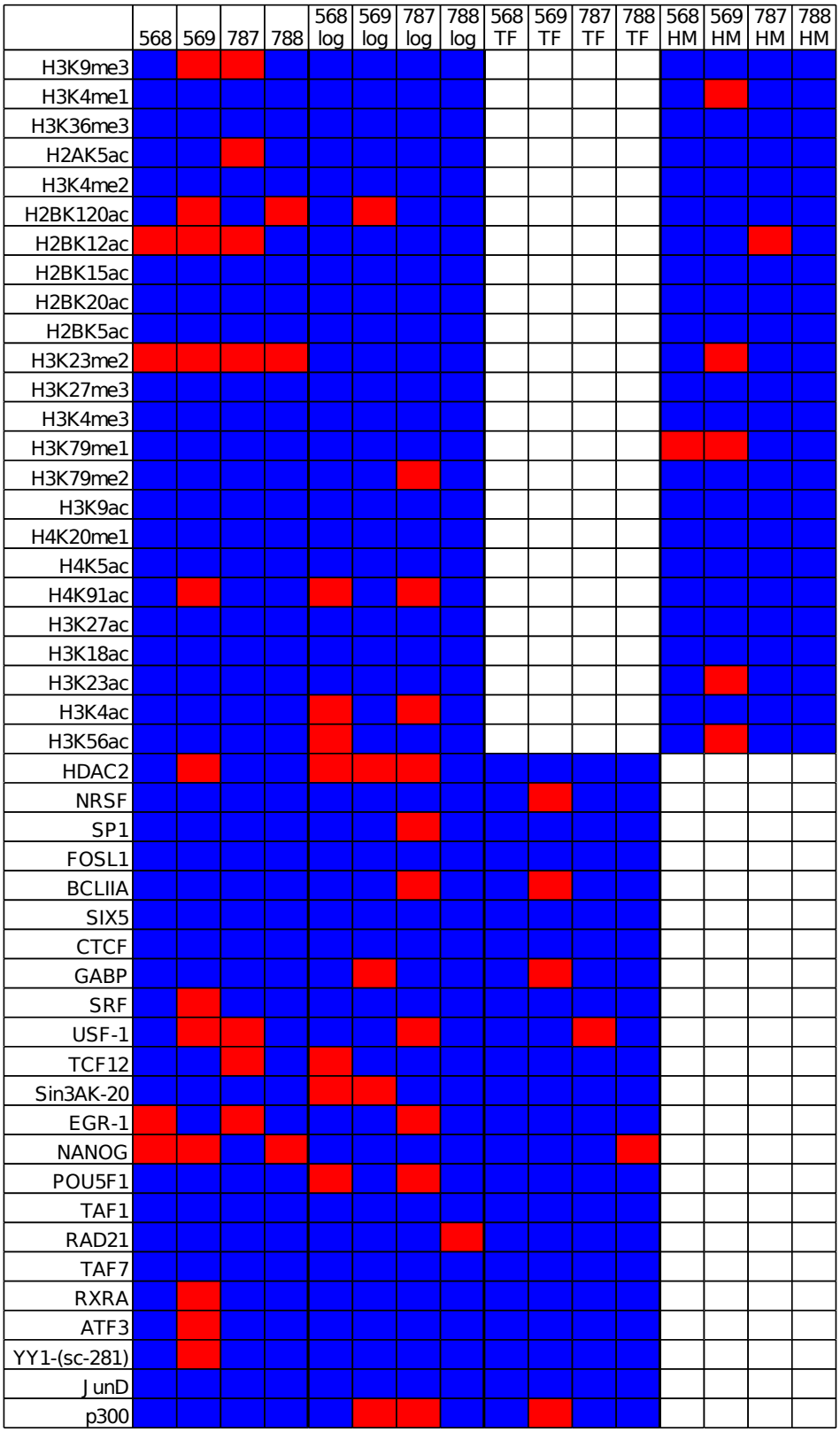


Figure 3.7: Heatmap showing the LASSO models for the four RNA polymerase II datasets. Sections in blue indicate the marker was included in the optimal model: sections in red indicate the marker was left out of the optimal model. The four RNA polymerase II datasets were abbreviated to the last three numbers of their accession code. TF: Transcription factor-only model. HM: Histone modification-only model.

out in more than one RNA polymerase II dataset for the transcription factor-only sub-set of data. Only a single predictor, H3K79Me1, is selected out from two models for the histone modification-only sub-set of data. For the non-logged and logged sub-sets, there are many more predictors that are selected out of multiple models. This consistency, along with more predictors being selected out of the non-logged and logged models, indicates that the models that encompass all of the data are more robust and reliable than the histone modification-only and transcription factor-only models. This, coupled with the better R^2 values across both of the whole-data models and compared to the split models, increases the likelihood of this being correct. For the non-logged models, there are four occurrences where a predictor is selected out of two models: H3K9Me3, H2BK120ac, USF-1 and EGR-1. There are two occurrences where a predictors is selected out of three models (H2BK12ac and Nanog) and a single case where a predictor is selected out of all four models (H3K23Me2).

Particularly of note here are EGR-1 and Nanog. EGR-1 (Early Growth Response protein 1) is a regulator of, and required for, differentiation. EGR-1's absence from the model is interesting, but not surprising, as the genes it targets for transcription promote differentiation and mitosis, both of which are suppressed in embryonic stem cells. Nanog, a protein that is essential for self-renewal and maintenance of undifferentiation of embryonic stem cells is also missing. The absence of the Nanog binding data from the majority of the non-logged-data models is interesting because it is vital for the establishment of pluripotency. Since Nanog is statistically redundant with the information contained in the other datasets, either its effect is too small to see using the broad metric of the enrichment values or its effect is already accounted for by the rest of the data. As Nanog is only needed to *establish* pluripotency, not *maintain* it this could also explain this result.[140] It is possible that the effect Nanog has on the other markers in the model makes the data for the gene statistically redundant in the model.

For the logged models, there are five occurrences where predictors are selected out of two models: H4K91ac, H3K4ac, Sin3AK-20, Pou5F1 and p300. There was also a single case where a predictor was selected out of three models: HDAC2. There are two interesting proteins that are not included in multiple of the logged-data models: Pou5F1 and Nanog. Pou5F1 is an essential protein for the maintenance of pluripotency. As with Nanog for the non-logged models, presumably the effect it has on RNA polymerase II binding is either lost in the broad study or its effects are already accounted for by the other datasets. HDAC2 is a protein responsible for binding the acetyl group to histones. HDAC2 is likely not included in the model in these cases as the information in the HDAC2 dataset is contained in the datasets of the acetylated histones.

3.3.3 Discussion

The comparison of the H1, H9 and IMR90 human embryonic stem cell lines clearly shows that it is not valid to use datasets from different cell lines with each other. While it may be possible for some cell lines, ChIP-seq data is noisy already without potentially introducing biological noise on top of the noise introduced by the techniques used.

Overall, the simple enrichment value metric, coupled with LASSO regression, results in accurate models for predicting the binding of RNA polymerase II. The use of LASSO regression to eliminate predictors from the models is effective and efficient. The best performing models use all of the datasets and the log of the enrichment values. This contrasts with the previous work where the optimal models were generated when only using histone modifications. This is likely due to having a much larger amount of data available for the cell line. While there is a large amount of statistical redundancy between the histone modification and transcription factor data, each individually does contain information the other does not. This is reinforced by the findings of Cheng and Gerstein, who found that the addition of histone modification data to a transcription factor-only model or the addition or transcription factor data to a histone modification-only model improved the amount of variance accounted for by the model by about 10%.[76] A difference was also evident between the human data and the mouse macrophage data; in the human data, the transcription factor models were better than the histone modification only models, whilst the reverse was true for the mouse macrophage models. Using both the transcription factor and histone modification data not only results in better models for the human embryonic stem cell data, but also produced models which eliminated the most predictors. The comparison of the number of eliminated predictors in the all-data model to the models using only transcription factor or histone modifications shows that there is a redundancy between the information in the transcription factor data and the histone modification data. This is in agreement with previous work by Cheng & Gerstein where it was also found that transcription factor binding and histone modification data were statistically redundant in mouse embryonic stem cells.[75] This redundancy well less clear in the results of the previous chapter. For the larger dataset of the human embryonic stem cell data, the redundancy between the transcription factor and histone modification data is more obvious but less abundant, with many more datasets being eliminated from the final models.

As in the previous chapter, a direct comparison of the results between the models developed in this work and those of previous work by Cheng and Gerstein[105] and Ouyang *et al*[74] is not entirely relevant due to the vastly different objectives of the methods. A comparison is

still interesting however. Cheng and Gerstein were able to explain 72% of the variance in their response variable using support vector regression on 12 transcription factors and 7 histone modifications. Ouyang *et al* we able to explain 70% of the variance in their response variable using principle component analysis on 12 transcription factors. Here, we explained 80% of the variance in the RNA polymerase II binding data using LASSO regression on 23 transcription factors and 24 histone modifications. While the model produced here does explain more variance than the previous methods, it also has far more data available for training the models. While having more predictors almost always leads to better models, simply due to having more data to optimise on, the LASSO regression eliminates any predictors that add no extra information to the model, avoiding this over-fitting. While the purpose of the models differ, both the results here and the results of Cheng and Gerstein highlight the statistical redundancy between the transcription factor and histone modification data.

While the models themselves are good and provide a clear picture as to the general state of redundancy between the transcription factors and histone modifications, the more fine-grained picture of the mechanism of RNA polymerase II binding is less clear. Within the sub-sets of data (all data non-logged, all data logged, transcription factor-only, histone modification-only) there is little consensus between the RNA polymerase II datasets. The non-log and log models both have similar levels of corroboration between the RNA polymerase II datasets. There is, however, a large amount of variation between the sub-sets of data, with the transcription factor-only and histone modification-only datasets having very little consensus between or within themselves. This is most likely because the range of transcription factors and histone modifications is fairly broad and all of the datasets have at least a small amount of unique information. The increased number of predictors eliminated from the logged and non-logged models is a strong indication of the redundancy between the transcription factor and histone modification datasets. There is, however, little consensus in which predictors are eliminated from these models. This is potentially due to the redundancy between the predictors and the method in which the models are constructed. Predictors are eliminated in various models; this is potentially due to the modelling taking alternative “kinks” in the path to the final model. The selection of one predictor at a certain point will result in the elimination of others that have redundancy with it, or combinations of predictors the model already contains.

Further development of this method would involve using interaction terms in the model. The result would contain a multiplication of every combination of all the predictors, making it significantly more complicated, and also statistically impossible on smaller sets of data. The large number of genes in our dataset makes the inclusion of the interaction terms possible while the

LASSO regression method allows the elimination of unnecessary predictors and avoids over-fitting. The methods used here are very broad. The enrichment metric used is very simple and is measured over a section of DNA that is both large, relative to the size of transcription factor binding sites, but also small compared to the potential spread of a histone modification. The enrichment value has a huge potential for refinement, from the exclusion of genes and accounting for the GC-content in the background calculation, to adding a term for long-range interactions. The current method of using all of the genes as data points in generating the linear models also means that the model will only be true for the general case and will be less good for specific cases. Combining interaction terms with principle component analysis could highlight some interesting combinations of genes responsible for different functions in the cell.

None of the current methods, including the one developed here, deal with the difficult problem of enhancer sites. The remote location of enhancer sites from the gene or genes they regulate causes a massive problem in the prediction of gene regulatory mechanisms. It is not currently possible to predict the location of enhancer sites not is it possible to identify what genes they have an influence on. While the method of Ouyang *et al* accounts in some way for distal transcription factor binding, it assumes the further from the genes transcription start site the binding is, the less influence it has on the expression of the gene. This is not true in the cases of enhancers. One of the best ways to improve these methods will be to work on encompassing these remote, but highly influential, enhancer sites into the models.

Chapter 4

Hot-Spot Prediction in Protein-Protein Interfaces

4.1 Introduction

Understanding, and being able to predict, protein-protein interactions will be essential for the complete understanding of intra- and inter-cellular function, as well as elucidating the large protein-interaction networks that are present in every cell. All proteins are involved in a protein-protein interaction at some point between their translation and their degradation. Being able to predict and understand how protein-protein interactions occur will add vastly to the knowledge of how cells function as it will allow a better understanding of their precise function, and movements and actions within a cell.

4.1.1 Properties of the Interface

Size and Shape

The size of the interface involved in protein-protein binding varies considerably depending on the type of the interaction and the number of proteins involved. On average 12% of the solvent accessible surface area of each monomer in a dimer is involved in the protein-protein interface, rising to 17% of each monomer's surface for trimers and 20% of the surface area for each monomer for tetramers.[188] The binding of proteins leads to an average loss of 5-20% ($\sim$780Å^2) in solvent accessible surface area, 12% of the solvent accessible surface area of each monomer in a dimer is involved in the protein-protein interface, rising to 17% of each monomer's surface for trimers and 20% of the surface area for each monomer for tetramers.[188] The amount of solvent accessible surface area lost when two proteins bind is related to the molecular weight of the proteins involved.[189] Larger losses of solvent exposed surface area leading to stronger interactions.[189;190] The variability in the amount of solvent accessible surface area lost is large, with some complexes, such as the superoxide dismutase dimer, losing as little as 670Å^2 SASA per protein monomer (Figure 4.1), and other complexes, such as catalase forming a dimer, losing as much as 10,570Å^2 SASA per protein monomer (Figure 4.1).[191] While the size of the interface may vary between complexes, the shape often does not. Both obligate (proteins that are always bound together) and non-obligate (proteins that are bind transiently and are not always bound) protein-protein interfaces have been shown to be largely flat,[189] although non-obligate complex interfaces are less flat than obligates. Antibody-antigen and enzyme-inhibitor complexes are likely to bind in the largest cleft on this surface.[192]

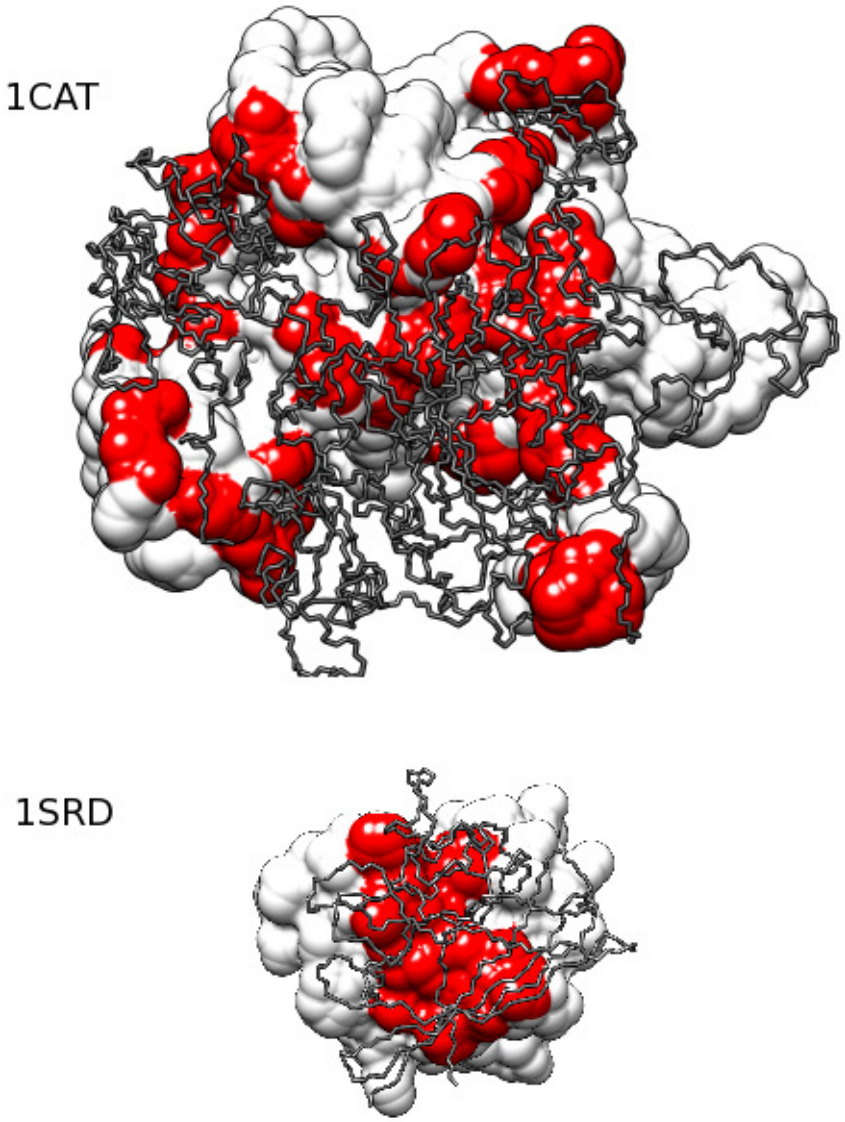


Figure 4.1: A contrast between interface sizes for protein-protein interactions. The catalase dimer (top, PDB ID 4CAT) loses 10,570Å^2 of solvent accessible surface area when the two protein chains form an interactions. The superoxide dismutase dimer (bottom, PDB ID 1SRD) in comparison loses only 670Å^2 of solvent accessible surface area when the two protein chains interact. Overlapping van der Waals radii between the two chains are shown in red .[191] Images generated using the UCSF Chimera package .[193]

Hydrophobicity

The hydrophobic effect drives a large number of the binding events in the cell and is a major driving force in the formation of the phospholipid bilayer that surrounds cells. When a hydrophobic molecule is introduced to bulk-water the polar water molecules re-arrange themselves to present as little polar surface to the molecule as possible, forming an organised solvation shell around the molecule. When more than one hydrophobic molecule is present in the system they cluster together so that as little surface as possible is exposed to the disorganised bulk water. This minimises the number of water molecules in the organised solvation-shell which leads to a rise in the entropy of the system. See figure 4.2 for a simple explanation of this. Hydrophobicity is a significant influence on the characteristic of protein-protein inter-

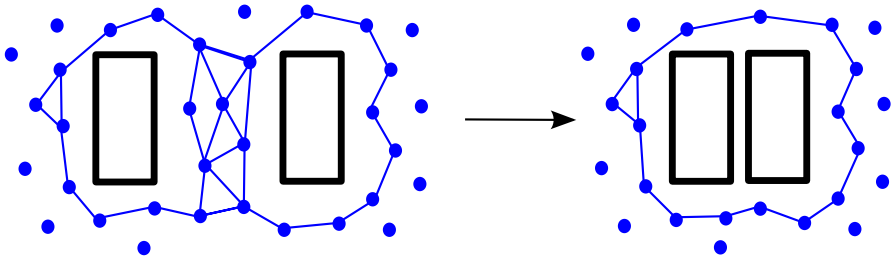


Figure 4.2: When a hydrophobic molecule is in water, the water molecules form a cage around it. This cage cause a reduction in the entropy of the system. When two molecules move together, the cage that is formed around them is needs fewer water molecules than the two individual cages, this raising the entropy of the system again. Image modified from Molecular Modelling: Principles and Applications, Leach, 2001 [194]

faces. [195] The characteristics of a protein-protein interface vary significantly depending on the proteins involved and hydrophobicity has a large influence on this. Interfaces between two identical chains are homo-oligomers and are often obligate complexes, which are only conformationally stable when bound to each other. Interfaces between non-identical chains are hetero-oligomers and are commonly non-obligates. Non-obligate protein-protein interactions occur between proteins that are in the same location within the cell; their binding can be triggered by a change in conditions, the association of a small molecule that induces binding, or binding and un-binding can occur continuously. [196] Due to the solvent-exposed nature of the unbound proteins, the domains involved in binding must be independently stable in solution. The independently-stable nature of non-obligate proteins contrasts with obligate binding proteins, whose interface surfaces are often conformationally unstable in solution. [188] The surface of the protein-protein interface in obligate complexes is generally hydrophobic and provides the driving force for the proteins to form a complex. When solvated the independent stability of the non-obligate complexes means that the elucidation of crystal structures is often much easier

for the unbound proteins than it is for bound complexes which are often only transiently stable. The opposite is true for obligate complexes as the equilibrium of the bound and unbound complex is so biased towards the bound complex that it is hard to isolate unbound proteins. The difficulty in generating crystal structures of non-obligate complexes often means that the structures of the component proteins are obtained singly and must have their interacting surfaces and binding conformations and surfaces predicted.

Amino acid composition

Protein-protein interfaces are generally more hydrophobic than the rest of the protein surface.[191] The residues present in a protein interface are similar to those that occur between domains within a protein,[188] with a bias towards aromatic and non-polar residues, and away from small polar residues. The level of hydrophobicity is very interface-dependant. Obligate interfaces on average have 65% non-polar, 22% polar and 13% charged residues,[191] while non-obligate interfaces have, on average, 55% non-polar, 25% polar and 20% charged residues .[190] This difference in hydrophobicity reflects the need for non-obligate protein interfaces to be independently stable in solution. The number of non-polar residues on the surface of an interface, and over the whole of an interacting protein, is no different to that of a small globular protein that does not bind any proteins .[197;198] This means that it is not possible to predict whether a protein will interact or to identify a binding region simply by looking at the *properties* of the residues that make up the surface of the protein.

4.1.2 Hot-Spots

The amount of energy that each residue involved in the protein-protein interface contributes to the binding is not equal, with some residues contributing significantly more than others. It was discovered that the mutation of two tryptophan residues to alanine resulted in the loss of affinity of human growth hormone to human growth hormone binding protein .[199] Hot-spot residues are defined as residues whose mutation to alanine causes a severe (>2 kcal/mol as quantitatively measures and defined by Bogan and Thorn, 1998) loss in binding free energy for the complex.[200]

“O”-rings

Analysis of alanine scanning mutagenesis data has shown that a number of residues around the hot-spot residues form “O”-rings. An “O”-ring is a sheath of amino-acid side chains that

exclude bulk solvent from the interaction formed by the hot-spot residue and the interacting protein.[200] It has been estimated that this brings about a doubling of the hydrophobic effect on the residues inside the "O"-ring.[201] Inside "O"-rings, hot-spot residues were found to be enriched by a frequency of over 10%) in tryptophan, arginine and tyrosine, whilst generally lacking (frequency in hot-spots under 3%) in leucine, methionine, serine, threonine and valine residues. There is, however, no preference for any single property, such as charged residues or hydrophobic residues, in the "O"-ring or the hot-spots.[200] It has been postulated that these residues are more frequent at hot-spots due to their ability to form multiple types of interaction, such as the ability of tryptophan to be involved in aromatic π-interactions, hydrogen bonds (as a donor) and as a large aromatic surface .[200] Several studies have highlighted the conservation of the enriched residues within protein families.[202–204] Multiple hot-spots can occur in a single interface[204] and they generally form clusters towards the center of the interface.[203] Hot-spots tend to couple with hot-spots from the interface surface of the binding protein.[205]

4.1.3 Experimental Analysis of Protein-Protein Interactions

Experimental prediction of hot-spots

Experimentally elucidating hot-spot residues in protein-protein interfaces currently involves a difficult and expensive process of alanine-scanning mutagenesis. Alanine-scanning mutagenesis is involves replacing residues in one of the two interacting proteins with alanine. This replacement leads to a loss of the side-chain of the original amino-acid and hence the loss of any interactions it made with the partner protein. The change in binding energy between the two proteins is then taken as the energy contribution that the mutated residue gives to the protein-protein interaction .[206] Although this is an effective method for establishing the binding energy contribution from a side-chain, it does have problems. If the mutation of the amino-acid effects the 3-dimensional structure of the protein, it can lead to changes in the interface surface. These changes can cause the two proteins to stop their close interacting, artificially inflating the calculated binding energy for the amino-acid.

4.1.4 Computational Approaches

Docking

Predicting the docking of two proteins is a complicated task that cannot be solved trivially. The nearly infinite number of possible orientations the two proteins can potentially be in with each

other means that a brute force approach is not viable to find the lowest energy conformation. More complex methods are needed that use information about the proteins and about their interfaces, thereby constraining the number of possible orientations for the two proteins. To constrain the number of possible orientations we need to understand the amino-acid composition of the surfaces, how they may possibly interact and any distinctive features they may contain.

Protein-Ligand Interaction Prediction

In predicting protein-ligand, and protein-protein, interactions there are generally two stages. First, large numbers of orientations of the two molecules to be docked are sampled and a set of potential "correct" orientations is gathered. The main challenge in this step is in developing an approach to this problem that has this "correct" pose in the final set of orientations whilst not being too computationally expensive. Sampling every possible orientation would be impossible due to the infinite number of different poses that are possible. The second step is to identify the "correct" pose from the subset of the possible poses obtained from the first phase.

Autodock[207] is one of the most cited docking programs available and has been through numerous iterations. Autodock uses genetic algorithms, based on evolutionary principles, to generate the final set of docking orientations. Genetic algorithms are based on evolutionary principles, in this case the arrangement of the two molecules is equivalent to a gene and the coordinates of the atoms for each molecule is the phenotype. It uses the AMBER force field[208] to calculate the interaction energy between the two molecules then assigns a "fitness" value based on the energy. The "fitness" values determines the likelihood whether the orientation will "survive" or "die". Variation can be introduced as "mutations" or mimicking the crossover of genes.

4.1.5 Binding Site Detection

Identifying binding sites is a difficult problem and there have been numerous approaches to the task; geometry based, energy based and knowledge based methods. Being able to predict the location at which a ligand binds to a protein is an essential first task in structure based drug design.

Geometry Based Approaches

Geometry based binding site detection methods approach the problem of identifying ligand binding sites by analysing the 3-dimensional structure of the protein with the aim of detecting the pockets in the surface of the protein. It has been shown that often, but not always, the largest pocket in the surface of the protein is the ligand binding site.[209] Methods such as POCKET[210] uses 3Å probes passed along each of the x, y and z axis. Pockets are identified where a period where the probes don't overlap with the protein is surrounded by periods where the probe overlaps with protein atoms. This method is highly dependent on the original orientation of the protein on the axes. LIGSITE[211] and Pocket Finder[209] use a similar method but pass the 3Å radius probe along the x, y, and z axis and also along cubic diagonals. This reduces the effect that the original orientation has on the pockets that are detected.

Energy Based Approaches

Energy based approaches to binding site detection, such as GRID,[212] approach the problem by estimating the interaction energy between a probe and the surface of the protein. Q-SiteFinder[209] uses the GRID force field to calculate the energy of a van der Waals probe with the surface of the protein as it is passed along a grid. These interaction energies are then spatially clustered to produce pockets. Whilst the geometry based methods tend to identify the largest pocket as the ligand binding site, Q-SiteFinder is often able to identify the ligand binding pockets when this isn't the case.[209]

Knowledge Based Approaches

Knowledge based binding site detection methods approach the problem using biological information not necessarily associated with the structure to identify potential binding sites. Some methods, such as ConSurf,[213] identify the ligand binding sites using sequence conservation. This method works well for enzymes as the active site is often highly conserved between proteins that bind the same ligand. If there are similar proteins with high-resolution structures, determined by NMR or X-ray spectroscopy, it is possible to perform homology modeling to identify conserved ligand binding sites.

All of these methods perform well when the structure of the ligand-bound protein is available. When the structure of the protein changes significantly when the ligand binds it is much harder to identify the correct binding site from the unbound structure. None of these methods perform well when trying to identify the ligand binding site is located in the interface between

two proteins of a complex. The huge number of potential ligands can also pose a problem if the ligand the protein binds is unknown.

4.1.6 Hot-spot prediction

The difficulty in predicting the hot-spot residues in a protein-protein interface is due to the lack of defining features which would otherwise help to determine whether the target patch of protein surface is involved in a protein-protein interface. Recently, it has been established that desolvation and conservation are two of the most important properties of a protein-protein interface. Cleft size and electrostatics, which are important when attempting to predict protein-ligand interactions, are much less important for protein-protein interactions. The function of the complex determines the effect each property has. For example, in enzyme-inhibitor complexes, there is a larger influence from cleft size and electrostatics whereas only desolvation is indicative of antibody-antigen interactions.[214]

Kortemme and Baker were one of the first to produce a method for predicting hot-spot residues in 2002.[215] They developed a computational alanine scanning method using free energy calculations. Their free energy calculation includes terms for polar interactions, shape-complementarity and a solvent interaction term which penalises interactions between hydrophobic residues and water. The change in binding free energy was calculated when each residue of an interface was mutated to alanine and any residue with a change in binding free energy above 1kcal/mol was identifies as a hot-spot residue.[215]

Li *et al* have worked to produce a hot-spot prediction method based on the number of atoms of a residue involved in sidechain-sidechain contacts; the difference between the number of atoms involved in favourable and unfavourable contacts and the difference between the number of favourable contacts and unfavourable contacts.[201] Hot-spot residues frequently have more atoms involved in sidechain-sidechain contacts, with a larger number of atoms involved in favourable contacts and also have a larger total number of favourable contacts.[201] Non-hot-spot residues have fewer atoms involved in sidechain-sidechain contacts, more atoms involved in unfavourable contacts and more unfavourable contacts.[201] When tested on homodimers, a method based on predicting hot-spot residues using these factors performed similarly to the energy based FOLDEF method,[216] but better than the structure based method PP_SITE[217] and the alanine scanning based method of Kortemme *et al*.[215] A problem with using the method proposed by Li *et al*[201] is that it fails to distinguish between hetero-dimeric complexes and crystal-packing artifacts produced by X-ray crystalography.[201]

The majority of hot-spot prediction methods require a 3-dimensional structure of the target protein in order to analyse the shape and structure of the surface. A novel approach has been taken by Grosdidier and Fernandez-Recio[218] using Normalised Interface Propensity (NIP) values. NIP values are generated for each residue in the protein from Free Fourier Transform-based docking methods.[219] Grosdidier and Fernandez-Recio ensured a robust NIP value by use of low-energy orientations for each of the training set proteins used to generate the NIP values. This enables the NIP-values to still function accurately in situations where larger conformational changes occurred and when no near-native structures were generated in the docking stage.[218] In the author's comparisons with well established methods such as Robetta[220] and Foldef[216] the method performs comparably, although producing a substantially lower sensitivity compared to Robetta; this is made up for by the flexibility of allowing the analysis of non-modeled proteins. Ofran and Rost[221] have developed another non-3D method, called ISIS, which is a neural network trained on all the interface residues in the Protein Data Bank.[222] ISIS produces low accuracy, low sensitivity and low specificity, because it was trained on a large selection of hot-spot and non-hot-spot interface residues and is based on sequence alone. This is likely because the amino acid sequence alone does not give information about what residues are in close proximity in the 3-dimensional structure of the protein.[221]

4.1.7 InterbasePro

InterbasePro is a database of predicted protein-biomolecule interaction energies. These predicted energies are calculated using MultiDock and optimised for DNA/RNA-protein interactions.[223] MultiDock is a molecular mechanics-based algorithm originally designed for calculating stable conformations in protein-protein interfaces but has been used for the calculation of the energy changes in the substitution of one amino acid to another in a protein-protein interface.[224]

MultiDock uses iterative mean field optimisation; a repeated two-step approach to calculating the interaction energy between two amino acid residues in a protein-protein interface. Firstly, each amino-acid side chain in the interface region has a set of conformational alternatives, rotamers, based on a library of possible rotamers.[225] Hydrogen atoms, not in the library, are then added to the rotamers. Iterative mean field optimisation is then used to calculate an energy for each possible rotamer and ranked by probability of occurrence. The next step is the use of the top-ranked rotamers in a rigid body energy minimisation to reduce any unfavourable interactions between the rotamer, the rotamers neighbouring amino acids and the amino acids

on the partner protein. These two steps are then repeated until no more improvement can be made for the intermolecular interaction energy.[223] The molecular mechanics potential energy function used by MultiDock is based on the AMBER force field.[208] The components of this force field and the molecular mechanics potential energy are described later.

Mean Field Optimisation

The mean field approach is used to calculate the potential energy for each rotamer.

$$E(i,k) = v(x_{ik}) + v(x_{ik}, x_{mc}) + \sum_{i=1, j\neq 1}^{N} \sum_{l=1}^{k_j} (CM(j,i)v(x_{ik}, x_{jl})) \tag{4.1}$$

For a protein comprised of N amino acids, each with side chain i and k_i possible side chain rotamers, the potential of mean force, $E(i,k)$, can be calculated. The first term, $v(x_{ik})$, is the internal energy of the rotamer where x_{ik} is the coordinates of the atoms of rotamer k of residue i. The second term, $v(x_{ik}, x_{mc})$, is the interaction energy between the rotamer and atoms of the main chain, where x_{mc} is the coordinates of the protein main chain atoms. The final term is the interaction energy between the rotamer and all the other rotamers of every other residue, weighted by their probabilities. CM is a conformational matrix of size N by $\max(k_i)$.

$$CM(i,k) = \frac{e^{-E(i,k)/RT}}{\sum_{k=1}^{K_i} e^{-E(i,k)/RT}} \tag{4.2}$$

With all the potentials acting on all K_i possible rotamers of the ith residue, the Boltzmann principle can be used to calculate the probability of rotamer k, as in equation 4.2, where R is the Boltzmann constant and T is the temperature. $CM(i,k)$ can then be used in equation 4.1 and the process repeats until the values within the conformational matrix converge at a minimum. This minimum is classed as when the root mean square of the deviation between the current and previous iteration, as calculated in equation 4.3 is less than 1x10^{-4}.

$$rms = \sqrt{\sum_{i=1}^{N} \sum_{k=1}^{K_i} (CM(i,k) - CM_{old}(i,k))^2} \tag{4.3}$$

Rigid Body Minimisation

The rigid body docking stage of MultiDock is to reduce any unfavourable interactions between the two binding proteins. For efficiency, only residues that have another β-carbon within 15Å of their β-carbon had their energy minimised. For glycine, the distance from the α-carbon is used.

The larger of the two proteins to be minimised is kept rigid and immobilised for the rigid body minimisation while the other protein is moved and rotated to find the minimum intermolecular energy. Steepest descent was used for the minimisation with a maximum rotation of 1° and maximum translation of 0.3Å, finishing when the energy decrease obtained by the successive minimisation steps falls below 1x10^{-6}kcal/mol.

Molecular Mechanics Potential Energy Function

Molecular mechanic potential energy functions are a combination of empirical force field parameters and Newtonian mechanics that, given the coordinates of the atoms in a system, can be used to calculate the potential energy of the system. The general equation for the potential energy function is as described in equation 4.4.

$$E = E_{elec} + E_{vdW} + E_{stretch} + E_{angle} + E_{torsion} \tag{4.4}$$

Where E is the potential energy function, E_{elec} is the energy from electrostatic effects, E_{vdW} is the energy from the van der Waals interactions, and $E_{stretch}$, E_{angle}, and $E_{torsion}$ are the energy contributions from bond stretching, changes in bond angles and the dihedral angles/torsions respectively.

Electrostatics

The electrostatics portion of the potential energy function is calculated using the Coulombic interaction energy between two atoms, as described in equation 4.5.

$$E_{elec} = 332 \left(\frac{q_i q_j}{r_{ij} \epsilon} \right) \tag{4.5}$$

q_i and q_j are the charges for atom i and j respectively. r_{ij} is the distance between i and j. ϵ is the dielectric constant and is used to account for the effect of highly charged environments. To have the resulting value in units of kcal/mol, a scaling factor of 332 is used.

van Der Waals

The van der Waals portion of the potential energy function is described using the Lennard-Jones potential (equation 4.6). This equation elegantly describes the manner in which at a distance, the van der Waals effect is mildly attractive. This attraction increases until their electron “clouds” begin to overlap, where upon the van der Waals effect becomes increasingly,

and rapidly, more repulsive. The first term of equation 4.6 describes the repulsive component, the second describes the attractive component.

$$E_{vdW} = \epsilon \left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \tag{4.6}$$

r is the distance between the two atoms, *epsilon* is the van der Waals well depth and σ is the distance at which the atoms start repulsing each other, or the collision diameter.

Bond Stretching

The bond stretching term is the deviation from normal in the length of the bond. A lot of energy is required to change the length of a bond and as such the deviations from the normal are very small. This means that it is possible to approximate the energy using Hooke's Law.

$$E_{stretch} = \frac{k}{2}(l - l_0)^2 \tag{4.7}$$

Where k is the propensity for stretching, l is the bond length, l_0 is the bond length where the energy is at the lowest. The term $(l - l_0)$ describes the deviation from the minimum energy.

Bond Angles

The bond angle term describes the energy involved in the angle between the atoms deviating from the optimal angle. As with the bond stretching term, the variations in this value are small and it can be approximated using Hooke's Law, equation 4.8.

$$E_{angle} = \frac{k}{2}(\theta - \theta_0)^2 \tag{4.8}$$

Where k is a constant describing the effects of changing the angle, θ is the angle, θ_0 is the angle where the energy is at a minimum and as such, $(\theta - \theta_0)$ describes the deviation of the angle from the minimum.

Torsion Angles

The torsion angles refers to the rotation about the bond. The profile of this energy is periodic in that it depends on the number of surrounding atoms and their locations. The energy of the angle depends on how well the atoms around the bond fit into "slots" between the surrounding

atoms.

$$E_{torsion} = \sum_{n=0}^{N} \frac{V_n}{2}[1 + \cos(n\omega - \lambda)] \tag{4.9}$$

Where n is the number of energy minima in a 360° turn, λ identifies where in the turn the minima occur and ω is the torsion angle. V_n describes the energy required to rotate the atoms between the minima.

Hydrogen Bonding

The number of hydrogen bonds formed by a complex has been shown to correlate strongly with the size of the interface[197] but, on average, protein-protein interfaces contain about ten hydrogen bonds.[195;197] The number of hydrogen bonds, as well as their strength, is dependent on the type of interaction. Enzyme-inhibitor complexes tend to form hydrogen bonds between their backbone atoms across the interface due to their close proximity and the manner in which they bind into a cleft, while antibody-antigen complexes tend to form more hydrogen bonds between side-chain groups.[195;197]

4.1.8 CAPRI

The Critical Assessment of Prediction of Interactions (CAPRI) was initiated to provide a blind test of current protein-protein interaction prediction methods with the aim of comparing the effectiveness of current methods as well as encouraging development of new methods. The test cases for the CAPRI tests were provided by experimental researchers who have elucidated the 3-dimensional structures of proteins whose bound structure was previously unknown. The aim is to predict the bound structure of the protein complex when given the structure of the unbound components without knowing any information about the target proteins before hand. CAPRI assesses all the residue-residue contacts between the two proteins as well as the residues that contribute to each side of the interface. For each of the submitted models the fractions of the native and non-native contacts in the predicted model, the root mean square deviation of the backbone atoms of the ligand protein and the mis-rotation and alignment of the ligand compared to the native conformation.[226–228]

4.1.9 Computational Hot-Spot Prediction Methods

There are a few methods currently available for predicting potential hot-spot residues using a variety of different methods. Foldef[216] is one of the earliest computational methods for predicting hot-spot residues and uses a full atomic description of the structure of the proteins along with empirical data obtained from protein engineering experiments to predict hot-spot residues. A full atomic description means that Foldef is able to predict the energy each residue contributes to the protein-protein interface by calculating all of the forces that are applied to and by each atom of the reside, and hence predict the influence the residue has on the interface. The correlation between the predicted changes in Gibbs free energy when a residue is mutated to alanine and the experimental values was 0.7, setting a high benchmark for future methods. Robetta[229] was also an early method for predicting hot-spot residues. Robetta uses a linear combination of the Lennard-Jones potential, a solvation term, hydrogen-bonding term and a term for rotamers. This method was able to identify 79% of hot-spots in their test dataset and is still one of the better methods available. There have been more recent methods that have been developed around an empirical model, such as the Hotpoint web-sever.[230;231] Hotpoint uses additional factors such as conservation, solvent accessible surface area and pair-wise residue potentials to predict hot-spot residues. It has 70% accuracy rate while retaining a precision of 64% which is impressive in this difficult field. A number of more recent methods have been focused on using machine learning to identify the best features to use for predicting hot-spot residues. MINERVA[232] uses a support vector machine based on features such as structure, sequence, molecular interactions and conservation along with a decision tree to establish the best combination of features. MINERVA performs well, with a sensitivity of 0.44, specificity of 0.9 and precision of 0.65. Xu *et al.* have developed a method using a semi-supervised boosting support vector machine and using Random-Forests to select the features and prevent the support vector machine from over-fitting.[233] This method has a good recall of 0.77 but sacrifices precision (0.46) and specificity (0.6) to attain this.

All of the hot-spot prediction methods train and validate their results using data from ASEdb and BIC. They also all suffer from the same problem in that there is a very little experimental alanine-scanning mutagenesis data available. Both ASEdb and BIC are outdated, with more data available in the literature than in these databases. Even including this additional data, due to the cost, difficulty and time involved in performing alanine-scanning mutagenesis experiments, there is still very little data available. The lack of data is especially problematic for any methods using machine learning as it limits the amount of features that can be added be-

fore statistical problems arise. It causes issues in the development of all methods though as it limits the amount of data that methods can be tested and trained on and the types of complex that methods can be tested against.

4.2 Method

4.2.1 Adaptation of InterbasePro

InterbasePro is an energy-based method for predicting the interaction energy associated with residues in the protein-protein interface. It contains interaction energy profiles for protein-protein, protein-DNA and protein-RNA complex structures present in the Protein Databank.[234] The interaction energies are calculated using a modified version of the protein-protein interface refinement tool, Multidock. InterbasePro considers the contribution to binding from the electrostatic and van der Waals interactions.[235] The iterative mean field optimisation used in MultiDock is not applied to InterbasePro as the conformations given in the PDB are used. Rigid body minimisation is still used on the PDB to reduce unfavourable interactions. These processes are repeated until the interaction energy, calculated using the AMBER force field[208] cannot be improved any further.[235]

The calculation for the prediction of the binding energy of a residue using InterbasePro is set out in Equation 4.10.

$$pE = aE_{vdw} + bE_{ele} \tag{4.10}$$

pE is the calculated contribution to binding the residue makes, a and b is a normalisation constants, E_{vdw} is the InterBasePro predicted contribution from van der Waals interactions and E_{ele} is the InterBasePro predicted contribution from electrostatics. The results from InterBasePro were fitted to the experimental energies using a least squares linear model, as provided by the statistical package R[112], to give the coefficients a and b which scale the energetic contributions to reduce the difference between experimental and calculated energies.

A desolvation factor was also added to InterbasePro in the form of the energy required for the transfer of an amino acid from solvent water to octanol. This is representative of the transition of an amino acid from the surface of a bound protein, where it is solvated in water, to a bound state where it is surrounded by other amino acids. The values for the octanol-water transfer are taken from Radzicka and Wolfenden[236] and are detailed in Table 4.1. The desolvation factor was combined with the InterbasePro electrostatic and van der Waals predicted values as set out in Equation 4.11.

$$pE_{tot} = aE_{vwd} + bE_{ele} + cE_{totaldesolv} \tag{4.11}$$

pE_{tot} is the predicted contribution to binding the residue makes to the protein-protein interaction. E_{vdw} and E_{ele} are the predicted contribution to binding from van der Waals interactions and electrostatic interactions respectively. $E_{totaldesolv}$ is the octanol-water transfer energy for the residue as taken from Radzicka and Wolfenden.[236] *a*, *b* and *c* are the coefficients provided by a linear model which scaled the predicted data to best fit to the experimental results, as calculated using the R statistical computing program.[112]

Amino Acid	Octanol-water transfer energy / kcal/mol
Leucine	1.76
Isoleucine	2.04
Valine	1.18
Phenylalanine	2.09
Methionine	1.32
Tryptophan	2.51
Alanine	0.52
Glycine	0.00
Tyrosine	1.63
Threonine	0.27
Serine	0.04
Histidine	0.95
Glutamine	-0.07
Lysine	0.08
Asparagine	-0.01
Glutamic Acid	-0.79
Arginine	-1.32
Cystiene	0.00
Proline	0.00

Table 4.1: Octanol-water transfer energies[236] used as solvation factors for InterbasePro.

4.2.2 Alanine Scanning Mutagenesis Dataset

Two datasets were used for comparing the InterBasePro predicted binding energies to experimentally derived binding energies. The change in free energy when a residue of a protein-protein interface is mutated from the wild-type to alanine is directly related to the dissociation constant, shown by Equation 4.12.

$$\Delta G = -RT \ln(K_d) \tag{4.12}$$

ΔG is the change in Gibbs free energy of the system, R is the universal gas constant, T is the temperature and K_d is the dissociation constant of the two binding proteins. The link between free energy and dissociation constant allows a comparison of the difference made

in the binding affinity of two proteins when a residue is mutated to alanine and the energies predicted by InterBasePro.

$$\Delta\Delta G = \Delta G_{(wt)} - \Delta G_{(mut)} \tag{4.13}$$

The change in binding energy (Equation 4.13) is calculated as the difference between the Gibbs free energy for the wild type complex ($\Delta G_{(wt)}$) and the mutated complex ($\Delta G_{(mut)}$) Two datasets of alanine scanning mutagenesis experimental data were compared to the InterBase-Pro predicted energies. Firstly, the data for protein-protein interactions were used from the Alanine Scanning Database (ASEdb).[237] A summary of this dataset is detailed in table 4.2.

Protein	Number of Mutations	ΔΔG Standard Deviation(kcal/mol)	Structure Resolution / Å
1A4Y	28	0.98	2.00
1AHW	8	1.42	3.00
1BRS	14	2.55	2.00
1BXI	28	1.41	2.05
1CBW	9	0.64	2.60
1DAN	65	0.65	2.00
1DFJ	9	1.66	2.50
1GC1	32	0.39	2.50
1JCK	10	0.86	3.50
1RHG	26	0.51	2.20
1VFB	29	0.99	1.80
2PTC	1	-	1.90
3HFM	16	2.2	3.00
Total	275	1.47	-

Table 4.2: A summary of the dataset used from the Alanine Scanning Experiment Database (Thorn & Bogan, 2001)[237]

The second dataset was mined from the literature using only data from the primary publication rather than databases such as ASEdb and the Binding Interface Database (BID,[238]), although some data in these sources were not available from primary sources (table 4.3). Where there is overlap between the data in ASEdb and the second dataset, the values have been re-calculated from the primary publication rather than re-used from ASEdb.

The ASEdb data set comprises of 6 Enzyme-Inhibitor complex, 5 immune complexes and 2 “other” complexes (1 toxin/receptor complex and 1 growth factor complex). The alternative data set comprises of 7 enzyme-inhibitor complexes, 7 immune complexes and 2 “other” complexes (both protein/receptor complexes).

Protein	Number of Mutations	ΔΔG std Dev(kcal/mol)	Structure Resolution/Å	Source
1A4Y	28	0.98	2.00	a
1BRS	16	2.37	2.00	b,c
1BXI	32	1.11	2.05	d
1CBW	10	0.56	2.60	e
1DAN	65	0.64	2.00	a
1DFJ	8	1.45	2.50	a
1DQJ	21	1.83	2.00	e
1FAK	39	0.82	2.10	e
1FCC	10	1.60	3.20	e
1GC1	30	0.44	2.50	f
1JCK	10	0.93	3.50	g
1JRH	17	0.94	2.80	e
1JTG	7	2.24	1.73	e
1KTZ	16	0.71	2.15	e
1VFB	29	0.99	1.80	a
3HFM	16	2.37	3.00	h
Total	354	0.65		

Table 4.3: The alternative dataset for the comparison of InterbasePro predicted data to alanine scanning mutagenesis data. a - (Thorn & Bogan, 2001),[237] b - (Schreiber & Fersht, 1995)[239],c - (Schreiber & Fersht, 1993)[240], d - (Wallis et al, 1998)[241], e - (Fischer et al., 2003)[238],f - (Ashkenazi et al, 1990)[242], g - (Leder et al, 1998)[243], h - (Rajpal, Taylor, & Kirsch, 1998).[244]

4.2.3 Predicting Pockets and Atom Contact Counts

Q-Sitefinder is a method for predicting binding pockets on a protein surface.[209] A grid with a point separation of 0.9Å is built around the protein. The non-bonded interaction energy is calculated using the GRID forcefield.[212] Any grid point which is able to accommodate a methyl probe without overlapping with any protein atoms and with an interaction energy below -1.4 kcal/mol is retained. These points are spatially-clustered so that no probe is further than 1Å (for a grid with a separation of 0.9Å) center-to-center from the nearest grid point in the same cluster. The interaction energy is then summed over the cluster and the total energy of the grid points comprising the pocket is used to rank the pocket.

Once pockets are found using Q-Sitefinder on both proteins involved in an interaction, each protein is assessed to identify residues of its binding partner within these predicted pockets. Since protein-protein interactions do not normally bind in the largest cleft on the surface of the proteins no discrimination was made based on the size of the pockets that were predicted. For each of the occupied pockets, the corresponding scaled predicted energy from InterbasePro was compared in two stages - first using the values for the energy, then, secondly, using the rank of the pocket and the rank of the predicted energy.

4.2.4 Atom Contact Data

For each residue of a pair of interacting proteins, its 'atomic interaction count' was calculated as the sum of the number of atoms within 6Å of each atom of the residue, as calculated in equation 4.14

$$C_i = \sum_{0 \cdots n} k_n \tag{4.14}$$

C_i is the total count for residue i, n is the atom in the current residue and k_n is the number of atoms from the opposing protein within 6Åfor the current atom (i.e., an atom count for all atoms within the first shell of van der Waals contacting atoms).

A pseudo-alanine scanning count was established by only counting the number of atoms from the partner protein that were within the threshold distance across the backbone atoms of the amino acid and the β -carbon, mimicking the mutation of the residue to alanine (Figure 4.3). This pseudo-alanine scanning count was then subtracted from the total count to produce a third metric corresponding to the number of interactions that are due to the R-group of the residue, ignoring any interactions that result from the backbone atoms. Proteins that have a loop buried into clefts of their partner protein are likely to have backbone atoms that come into relatively close contact with atoms on the partner, but between which no forces or influence are exerted. In the pseudo-alanine scanning count, glycine is ignored due to the fact it lacks a β -carbon. The dataset used for this work was the Protein Docking Benchmark 3.0.[245]

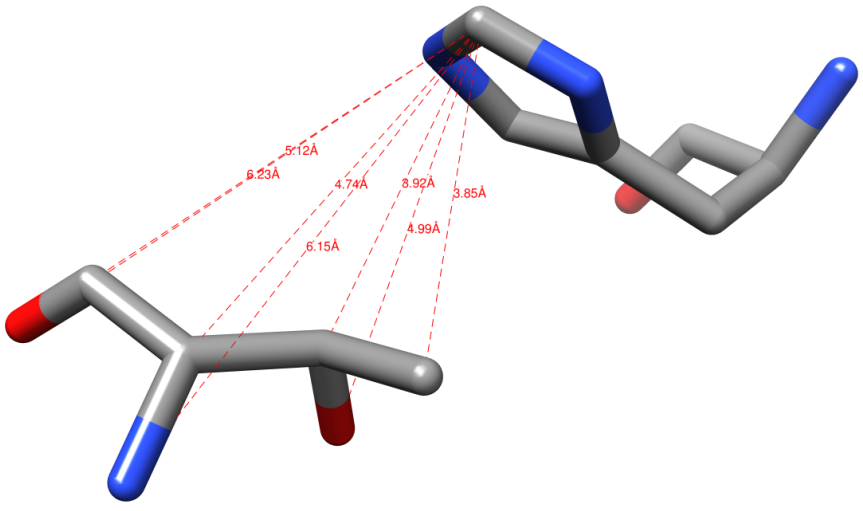


Figure 4.3: For each atom in a residue, any atoms of the opposing protein chain within 6Å are counted and regarded as having an interaction. A pseudo-alanine scan is performed by counting the number of interactions that are lost if the R-group was removed. In this case 4 of the possible 7 interactions are below the 6Å threshold.

4.3 Results & Discussion

4.3.1 Comparison of Calculations from InterBasePro with Experimental Results from ASEdb

Comparing the values of InterBasePro to the experimental data from ASEdb shows a significant difference between the predicted and experimental values. While for some complexes there is enough correlation to indicate that there is a relationship between the experimental energies and the computationally predicted ones, there are other complexes that indicate the opposite is true (Table 4.4). 1RHG and 1DAN are two complexes for which the correlation between the experimental and computational data is especially poor. 1RHG has a correlation coefficient of 0.02 and 0.05 for linear model scaled and unscaled correlations, respectively. 1DAN has a correlation coefficient of 0.15 and 0.17 for scaled and unscaled correlations respectively. For both of these complexes, the correlation coefficient implies that there is not a strong connection between the hot-spot residues as indicated by the experimental data and those that are predicted to have significant binding energies by InterBasePro. Two of the complexes where there does appear to be a connection between the experimental and computational binding energies are 1A4Y and 1VFB. 1A4Y has correlation co-efficients between the two methods of 0.47 and 0.48 for unscaled and scaled values respectively. 1VFB has correlation coefficients between the two methods of 0.50 and 0.51 for unscaled and scaled values respectively. These high correlation coefficients shows that there is indeed a connection between the alanine scanning experimental data and the values that were predicted by InterBasePro. The p-values for these complexes indicate that these results are likely to not be due to chance with values of $4x10^{-5}$ and $1x10^{-5}$ for 1A4Y and 1VFB respectively. Interestingly, the complexes where the correlation was strongest had the least improvement when they were scaled using the linear model. This is most likely because they were already close in value so the scaling had very little to shift in values, making its effect smaller than on the complexes where the correlation was poor, and there was greater scope for changing the predicted value to come more into line with the experimental results. The graph showing the experimental results from ASEdb and the predicted results from InterBasePro shows the correlation between the two methods (Figure 4.4). The overall correlation between the experimental data from ASEdb and the predicted results is 0.29 for both the unscaled and scaled values from InterBasePro. Scaling these results by the van der Waals and electrostatic energies from InterBasePro using a linear model had very little effect on the correlation between the experimental and computational

results. This low overall correlation is evident from the scattering of points in Figure 4.4. The graph also shows the number of true positive results (TP, points with high experimental energy and high predicted binding energy), true negative results (TN, points with low experimental and low predicted binding energies), false positive results (FP, points with low experimental but high predicted energies) and false negative results (FN, points with high experiential but low predicted energy). The cut-offs for a residue to be classed as a hotspot or not were 2 kcal/mol for the experimental energies, as used by Bogan and Thorn,[200] and 1 kcal/mol for the predicted energies. While there are similar numbers of true positive results and false positive results (Table 4.6) (top right and bottom right quadrants of the graph respectively), there are significantly more false positive results (top left quadrant of the graph) than either of these. There are, however, a very large number of true negative results (bottom left quadrant of the graph) where both the experimental and computational methods gave the residues low binding energies. This large number of true negatives is expected as non-hot-spot residues occur more frequently than hot-spot residues though being able to identify reliably what residues are not hot-spots can be useful.

PDBID	R^2	Scaled R^2	Number of Mutations	p Value (scaled)
1A4Y	0.47	0.48	28	0.00004
1AHW	0.06	0.33	8	0.13690
1BRS	0.42	0.42	14	0.01178
1BXI	0.36	0.50	28	0.00002
1CBW	0.00	0.37	9	0.08325
1DAN	0.15	0.17	65	0.00076
1DFJ	0.00	0.33	9	0.10340
1GC1	0.13	0.21	32	0.00889
1JCK	0.15	0.17	10	0.23280
1RHG	0.02	0.05	26	0.26050
1VFB	0.50	0.51	29	0.00001
3HFM	0.22	0.27	16	0.04746
All	0.29	0.29	274	<0.00001

Table 4.4: Correlation coefficients for the comparison of InterBasePro predicted changes in binding energy with the alanine scanning mutagenesis data from ASEdb.

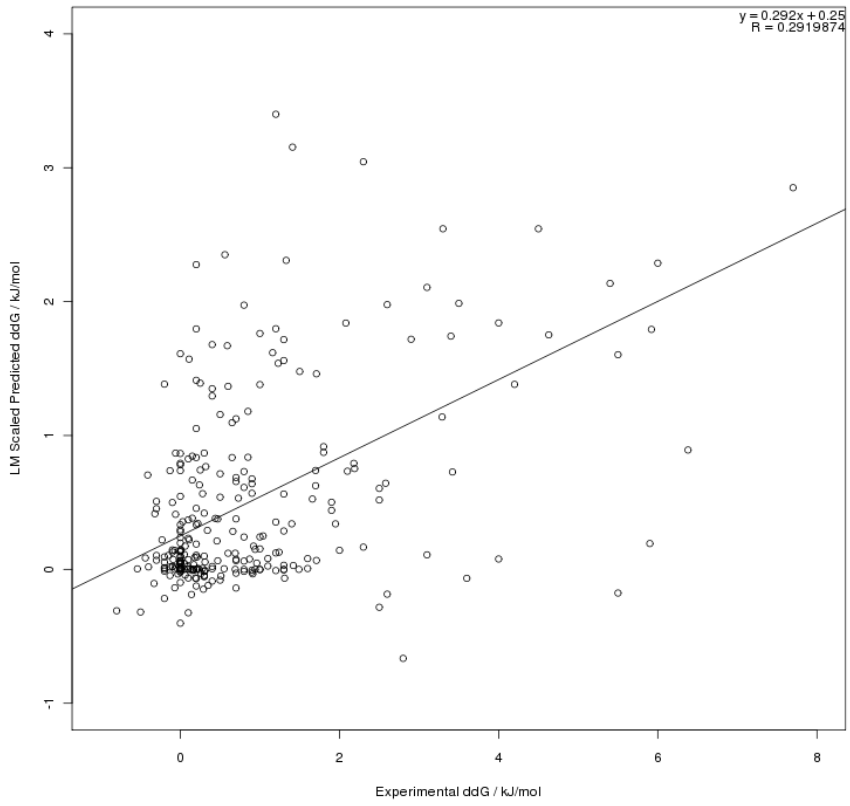


Figure 4.4: A comparison of the experimental alanine scanning ΔΔG from ASEdb and predicted ΔΔG from InterBasePro, scaled by electrostatics and van der Waals forces using a linear model.

4.3.2 Comparison of Calculations from InterBasePro with Experimental Results from the Second Dataset

As with the comparison of the ASEdb experimental results with the InterBasePro predicted energies, the correlation between the experimental data mined from the literature and the InterBasePro predicted energies are low (Table 4.5). The data mined from the literature is more comprehensive than the data from ASEdb. The correlation coefficient for the unscaled and scaled values is lower for the comparison of the mined-dataset with InterBasePro than that for the comparison on ASEdb and InterBasePro; 0.20 and 0.29 respectively.

PDB	R^2	Scaled R^2	Number of Mutations	p-value (scaled)
1A4Y	0.48	0.48	28	0.0002
1BRS	0.43	0.43	16	0.0200
1BXI	0.02	0.12	32	0.1600
1CBW	0.48	0.64	10	0.0300
1DAN	0.22	0.28	65	<0.0001
1DFJ	0.05	0.31	8	0.3900
1DQJ	0.15	0.23	21	0.0900
1FAK	0.34	0.36	39	0.0002
1FCC	0.38	0.66	10	0.0200
1GC1	0.10	0.14	30	0.1300
1JCK	0.11	0.21	10	0.4500
1JRH	0.12	0.17	17	0.2800
1JTG	0.11	0.48	7	0.2600
1KTZ	0.23	0.81	16	<0.00001
1VFB	0.54	0.55	29	<0.00001
3HFM	0.20	0.22	16	0.1900
ALL	0.20	0.20	354	<0.0001

Table 4.5: Correlation co-efficients for the comparison of InterBasePro predicted changes in binding energy with the experimental alanine scanning gathered from the literature.

For the mined-dataset, as with the ASEdb dataset, there are some complexes for which InterBasePro predicts the binding energy of the residues well, and some where it performs poorly. The correlation coefficients for 1CBW and 1FCC are good, especially when the electrostatic and van der Waals forces are scaled using a linear model; both reaching a correlation coefficient of over 0.6. There are several complexes, however, where the method performs less well. As in the previous graph, Figure 4.5 shows the extent of the poor correlation between the literature sourced dataset and InterBasePro. The number of true positives (top right quadrant), false positives (bottom right quadrant) and false negatives (top left quadrant) are about similar and the true negatives vastly out number all of these (Table 4.6).

The sensitivity and specificity values in Table 4.6 reinforce the idea that van der Waals and electrostatics interactions are not effective at predicting if a residue is a hotspot, with

InterBasePro having a sensitivity of about 50% for both data sets. The high specificities, 0.92 (ASEdb) and 0.88 (all data), however, show that InterBasePro is an very effective predictor of which residues are not hot-spot residues.

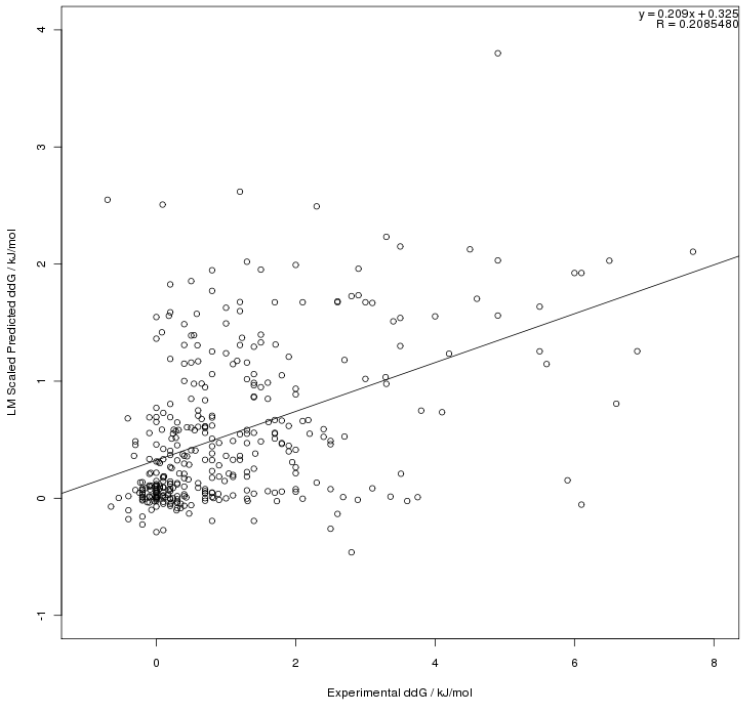


Figure 4.5: A comparison of the experimental alanine scanning ΔΔG mined from the literature and predicted ΔΔG from InterBasePro, scaled by electrostatics and van der Waals forces using a linear model.

	ASEdb	Alternative
True Positives	20	33
False Positives	31	46
False Negatives	17	34
True Negatives	205	240
Sensitivity	0.54	0.49
Specificity	0.92	0.88

Table 4.6: sensitivity and specificity for the linear model Scaled predicted values from InterBasePro compared with the ASEdb and alternative experimental data sets. The cut-off for being positive values for the experimental data is 2 kcal/mol and for the predicted values is 1 kcal/mol.

4.3.3 Adapting InterbasePro

A simple desolvation factor was introduced into InterbasePro to try to take into account the effects of desolvation on the stability of a protein-protein interaction. The ability to account

PDB	InterbasePro R^2	p-value	Octanol R^2	p value Octanol
1A4Y	0.4960	0.0001	0.0228	0.4430
1AHW	0.0041	0.0101	0.2790	0.1790
1BRS	0.4360	0.0005	0.0101	0.733
1BXI	0.3790	0.7380	0.1020	0.0971
1CBW	0.0170	0.0039	0.0173	0.7360
1DAN	0.1250	0.1430	0.0158	0.3190
1GC1	0.0701	0.8430	0.0000	0.9920
1JCK	0.0522	0.2640	0.1290	0.3090
1RHG	0.0515	0.2460	0.0649	0.2090
1VFB	0.4210	0.0001	0.0017	0.8310
3HFM	0.1210	0.1870	0.0009	0.9140
All	0.2280	$<2.20\text{x}10^{-16}$	0.0047	0.2650
Enzyme/Inhibitor	0.3010	$<2.2\text{x}10^{-16}$	0.0016	0.5480
Immune	0.0794	0.1070	0.0098	0.5790

Table 4.7: Correlation coefficients for the comparison of the predicted energies from InterbasePro and the experimental energies of ASEdb.[237] *InterbasePro* R^2 is the correlations comparing InterbasePro to the experimental data while *Octanol* R^2 refers to the data from InterbasePro with an extra octanol-water transfer energy component added to act as a solvation factor.

for the effect of solvation is important as it is a significant contributing factor in the formation of protein-protein interactions and the "O"-rings are thought to exclude the solvent from the hot-spot residues.[200] Using the calculated energy for the transition of an amino acid from water to octanol is relatively similar to the transition of an amino acid from a solvated, unbound state in water to a bound state in the mixed hydrophobic/hydrophilic environment of the protein interior.[246;247] Combining a simple scaled octanol-water solvation factor with the energies predicted by InterbasePro causes an overall decline in the correlation between the predicted and experimental energies (R^2 = 0.0047, table 4.7). The addition of a desolvation potential does have a positive effect on the correlation coefficients for some proteins, 1AHW and 1JCK in particular, but the p-values for these proteins show the results are not significant (p-value > 0.005). A more complex, atom-by-atom solvation potential would more likely be representative of the effect of desolvation of residues as they go from the unbound to the bound state. Such a method based on the work of Wesson and Eisenberg could be implemented[248]; a solvation factor based on the loss or gain of accessible surface area for each type of atom rather than, as used here, a single value used for each amino acid regardless of its exposure to the solvent would more accurately reflect the influence of solvation. InterbasePro is an energy based method that encompasses electrostatics and van der Waals terms in the prediction of binding energies for each residue in a protein. Comparing the predicted energies from InterbasePro to the experimental data from ASEdb in Table 2 shows poor correlation (R^2 = 0.228). Figure 4.6 and 4.7 show comparison of the experimental energies and the predicted energies for proteins in the ASEdb. There are a large number of residues with high experimental energies but with

low predicted energies. There are also a large number of residues with a low experimental energy but high predicted energy. There are very few residues where both the experimental energy and the predicted energy were high. This lack of correlation between the experimental and predicted energies could be due to the lack of a solvation factor being included in the prediction of the energy. A potential flaw in the experimental data is that the change in the binding energy recorded could be due to disruption of the 3-dimensional structure of the protein, potentially causing it to unfold or take on another conformation. If the mutation of a residue to alanine results in a conformational change in the protein it can lead to a large change in binding energy, regardless of whether the residue was involved in the protein-protein interaction or not. One approach could be to investigate the interactions of side chains with the side chains and backbones of the residues surrounding it on the same protein chain. The more interactions that are lost between the mutated residue and the backbone structure of the partner protein the more likely it is that the mutation disrupts the conformation of the protein.

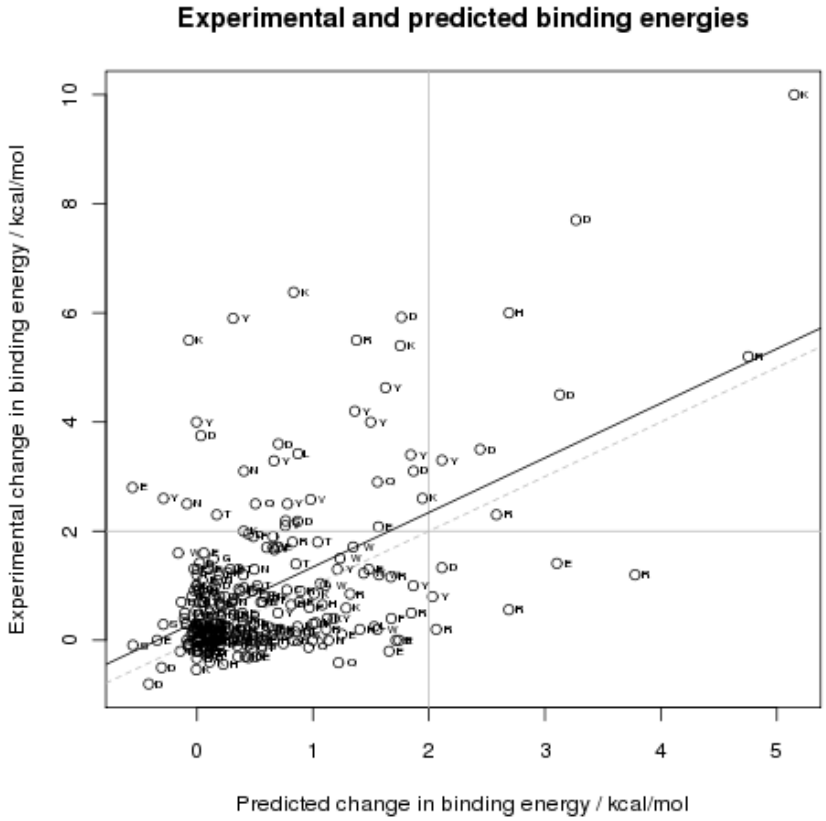


Figure 4.6: Comparison of the experimental alanine scanning mutagenesis data from ASEdb[237] and predicted energies calculated by InterbasePro. Points are labeled with their respective 1-letter amino-acid code.

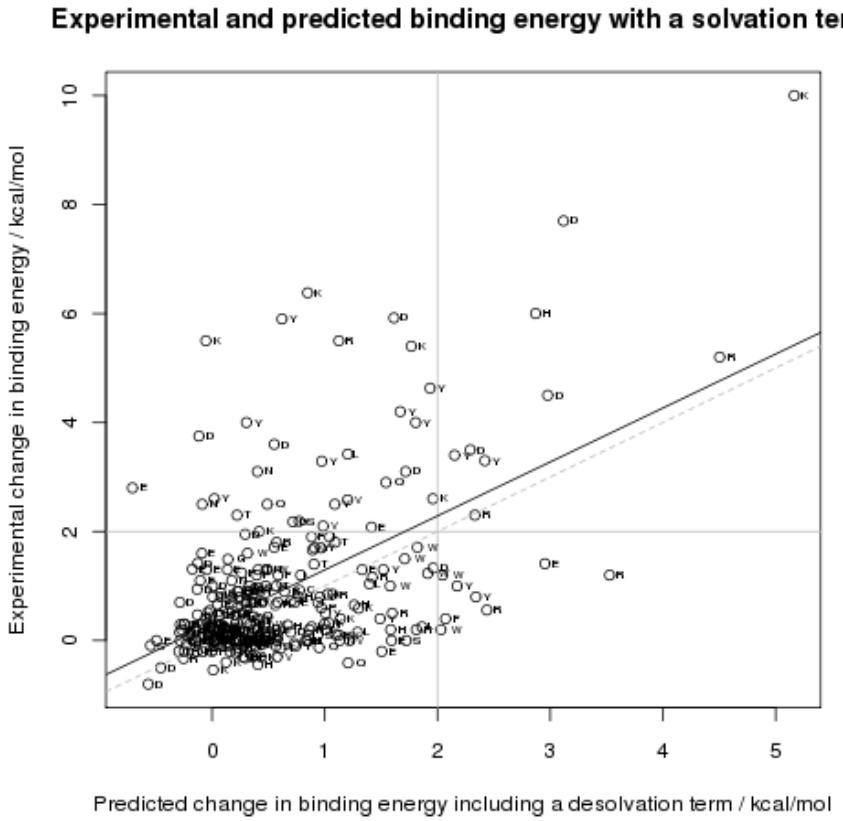


Figure 4.7: Comparison of the experimental alanine scanning mutagenesis data from ASEdb[237] and *unscaled* predicted energies calculated by InterbasePro with an added desolvation term based on the octanol-water transfer energy of amino-acids.[236] Points are labeled with their respective 1-letter amino-acid code.

4.3.4 Atom Contact Data

The atom contact data (Table 4.8) show a positive correlation between the number of interactions and the experimental change in free energy when the residue is mutated to alanine in alanine-scanning mutagenesis experiments. The pseudo-alanine scan count is show relative to the experimental energy change in Figure 4.8. The overall trend is that as the number of interactions made between a residue and the opposing protein increases, so the higher the change in experimental free energy when the residue is mutated to alanine. Three proteins

	Total	Total p-value	Diff. R^2	Diff. p-value
1A4Y	0.226	0.0026	0.158	0.0134
1AHW	0.432	0.0147	0.746	0.0001
1BRS	0.255	0.0045	0.433	0.0001
1BXI	0.239	0.1046	0.321	0.0546
1CBW	0.238	0.0401	0.191	0.0696
1DAN	0.044	0.1955	0.021	0.3744
1GC1	0.042	0.4616	0.012	0.6960
1JCK	0.371	0.0000	0.054	0.1287
1RHG	0.005	0.6899	0.000	0.9633
1VFB	0.408	0.0010	0.458	0.0004
3HFM	0.115	0.1801	0.199	0.1101
3HHR	0.344	$1.92\text{x}10^{-9}$	0.255	$5.35\text{x}10^{-7}$
All	0.081	$2.04\text{x}10^{-8}$	0.055	$4.25\text{x}10^{-6}$
Enzyme/Inhibitor	0.249	$4.41\text{x}10^{-4}$	0.254	$2.22\text{x}10^{-16}$
Immune	0.152	$2.60\text{x}10^{-4}$	0.161	$2.29\text{x}10^{-4}$
Without Outliers	0.236	$2.20\text{x}10^{-16}$	0.238	$2.20\text{x}10^{-16}$

Table 4.8: Correlation coefficients for the atom count data. Total is the correlation coefficient when regarding all the atoms in each residue. *Diff. R^2* and *Diff. p-value* are the difference correlation and the p-value for the difference between the number of interactions made between the whole of a residue and the partner protein and the number of interactions made by the backbone atoms and the β -carbon. The outliers removed in the "without outliers" data are 1RHG and 1DAN.

disobey this trend: 1RHG, 1GC1 and 1DAN. Removing these proteins from the calculation of the total Pearson's correlation coefficient calculation has a positive effect on the overall correlation due to the outlier nature of the results for these two proteins. Withdrawing these proteins from the calculation increases the total correlation coefficient from 0.081 to 0.236 and the pseudo-alanine scanning correlation coefficient from 0.055 to 0.238. The extremely low correlation coefficients for 1RHG (R^2=0.005, p-value=0.6899), 1GC1 (R^2=0.042, p-value=0.4616 and 1DAN (R^2=0.044, p-value=0.1955) are due to a few residues which have a massive number of contacts with the partner protein, far more than any other residue in the dataset, which also have a small experimental energy change.

Overall the correlation coefficients show some agreement between the number of interacting atoms and the energy changes produced by mutation to alanine. Some proteins, such

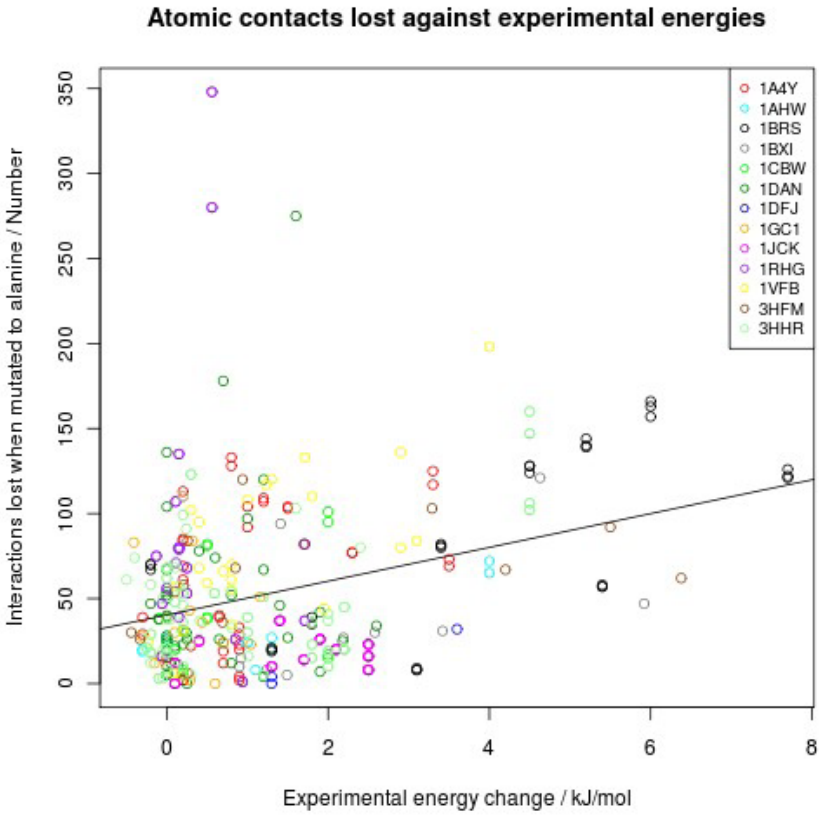


Figure 4.8: Comparison of the atom contact count lost when a residue is mutated to alanine and the experimental energies from ASEdb.[237] coloured by protein

as 1AHW, 1VFB, 1JCK and 3HHR, show significant correlation and p-values. While they are still weak correlations they show the potential in this very simple method. Within the proteins that for which the method functions well, there is a large variation in how effective the method is between the total counts and the pseudo-alanine scanning count. 1JCK has a reasonable correlation for the total count data, R^2=0.371, but performs poorly, R^2=0.054, for the pseudo-alanine scanning. This difference is due to the close proximity of the backbone of the two bonding chains.

Comparing the correlations for the data from enzyme-inhibitor complexes and the systems involved in immune response shows that this method is more reliable for enzyme-inhibitor complexes than immune complexes. This is most likely due to the weaker, more transient, and generally more distant binding that occurs in antibody-antigen binding.

4.4 Conclusions

All the results comparing the predicted results of InterBasePro with the experimental alanine scanning results of ASEdb, as well as the data mined from the literature, show that the ability of InterBasePro to predict which residues in a protein's surface are hot-spots is poor. The comparison of the predicted energies from InterBasePro, whether scaled using a linear model or not, had a correlation coefficient of 0.29. The comparison of the predicted energies of InterBasePro with experimental energies mined from the literature has a correlation coefficient of 0.2. While generally poor, InterBasePro performs well for some complexes. What the results also show is that InterBasePro is very effective at predicting which residues are not hot-spots. This could mean that, although InterBasePro alone would not be effective at predicting which residues are hot-spots in the general case, combined with further methods or factors, it could be used to predict hot-spots. The use of all the data together for the model is likely not the best way to approach the problem of predicting hot-spot residues. With extra information about the function or type of complex the target proteins are it may be possible to tailor the methods used to predict hot-spot residues.

Compared to Foldef and Robetta, both the fitted InterbasePro model and InterbasePro with a solvation term performed poorly. On data from ASEdb, Foldef obtained a correlation coefficient of 0.7.[216] This is far in excess of the 0.2 and 0.004 for the InterbasePro and InterbasePro with solvation term models, respectively. In comparison to Robetta,[220] again, both methods used here performed worse, with Robetta being able to identify a far higher proportion of the true-positive results than both InterbasePro and when InterbasePro is combined with a solvation term. The differences between the prediction accuracy of the InterbasePro-based methods and that of Foldef and Robetta is likely due the development process that InterbasePro has undergone. Originally, InterbasePro was designed for predicting protein-DNA interactions, despite the method it is based on, MultiDock, being designed for predicting protein-protein interactions. The weightings of the energy terms were changed to optimise the method for protein-DNA interactions, with more emphasis being placed on electrostatic effects. While InterbasePro is not specific to protein-DNA interactions, the results here show that it has lost the predictive power of its underlying method when re-applied to hot-spot prediction. In comparison, Foldef and Robetta were both designed purely for predicting hot-spots in protein-protein interactions. While the theory behind the solvation term used is logical, in actuality the mechanisms behind solvation are far more complex than the assumptions in using just these solvation terms. The decrease of the predictive power of the method when the solvation term

is used in conjunction with the InterbasePro terms is likely due to the compound effects of the design of InterbasePro with the simplistic assumptions made in the use of such a simple solvation term. Though disappointing, predicting hot-spot interactions is a difficult task and requires many rounds of iterative improvement on the methods used. Like InterbasePro, both Robetta and Foldef use computation descriptions of the atomic interactions between all the atoms of the residues of the protein, highlighting that InterbasePro could potentially be improved in it's ability to predict hot-spot residues. While the correlation between InterbasePro and the experimental results is low, it is positive. This slight correlation does mean that it could potential be combined with other metrics, such as the atom count method developed here, to improve on its results.

The electrostatic and van der Waals interactions from InterBasePro are not predictive for which residues are hot-spots but can be effective are predicting which residues are not hot-spots. This is in agreement with previous work done by Burgoyne and Jackson[214] which shows that the only surface property that is consistently predictive of a protein-protein binding region is the ease of cleft desolvation. Electrostatics, van der Waals, desolvation and surface conservation showed little predictive power for whether a region of a protein's surface was and interface region. A potential future direction that will enable InterBasePro to act as an effective hot-spot prediction method will be the introduction of a desolvation factor into the prediction of energies. An atom-by-atom desolvation factor will potentially enable InterBasePro, after filtering unlikely hot-spots residues, to filter and separate the true positive and false positive results.

A significant problem in this area is the lack of experimental alanine-scanning mutagenesis experimental data. Alanine-scanning mutagenesis is a complicated and time-consuming technique and, despite the wealth of information it provides, it is not often undertaken. Although ASEdb is a good resource for this data, it is slowly becoming out-dated and not all of the data appears to be available in the literature, having been uploaded by individual researchers. The alternative data set produced from literature searching is far from ideal, having significant overlap with the data from ASEdb, albeit with values calculated independently from the literature where needed, with neither data set being extensive. The small amount of data available will potentially lead to a very narrow view of hot-spot properties and complicate the training of any methods produced. More data would also give a better comparison between the different types on complex and allow a more tailored approach to the problem, with adjusted calculations for different complex types or functions.

As computing power increases in the future, it may well become possible to do full free-

energy calculations on protein-protein interactions. At the current time it is still very computationally expensive to do these calculations in full for large scale systems. As the available processing power increases it may well be possible to, given some assumptions to simplify calculations and potentially using rigid-body modeling, to reduce the computation time needed to a level where these complexes are able to be modeled.

Chapter 5

Summary

The initial chapter of this work establishes the use of a simple metric and linear regression to model the binding of RNA polymerase II. The simple metric, based on the enrichment of reads over the promoter and gene region compared to a background value, was used in all of the models in the first two chapters. Initially this method was applied to a set of 5 transcription factors and 4 histone modifications from mouse macrophage cells. The models produced showed a distinct redundancy between the histone modification and transcription factor datasets. The best model only used 3 histone modification datasets to model RNA polymerase II data and was able to explain 78% variance in the RNA polymerase II data. An investigation of the genes with the largest and smallest residues showed no over-representation in genes specific to macrophage cells. A comparison between the histone modification-only model and the next best model, which used 3 histone modifications and 3 transcription factors, was done to investigate the effect the transcription factors had on the models. By removing each transcription factor, and every combination of the transcription factors, and comparing the resulting model with the model using the 3 histone modifications and 3 transcription factors, it was possible to look at which genes were affected most by the transcription factor/s in question. Analysis of these most and least affected genes again showed no pattern in genes or over-representation in macrophage specific genes.

By validating this method on the small set of data for mouse macrophage cells it allows us to apply it to a larger dataset. A set of 47 ChIP-seq datasets from human embryonic stem cells from the ENCODE project was used in the subsequent investigation; the datasets were for 23 transcription factors and 24 histone modifications. Due to the large number of predictors available, standard linear regression was not possible. LASSO regression was used to eliminate predictors that contributed no extra information to the model. The resulting models varied greatly in the numbers of predictors that were used in the model for each RNA polymerase II dataset, with between 1 and 10 predictors being eliminated in each log-enrichment model. The log-enrichment models performed the best and were able to explain between 79-84% of the variance in the RNA polymerase II datasets. The statistical redundancy between the transcription factor data and the histone modification data was especially clear from these models. The histone modification-only and the transcription factor-only models eliminated between 0 and 5 predictors while the models that had all of the datasets available eliminated far more predictors, up to 12. Models were produced for the 4 RNA polymerase II datasets. There was little consensus between the predictors used for each of the models produced, most likely due to slight variations in the RNA polymerase II datasets leading to different predictors being incorporated into the model and ultimately a different set of predictors being used for the final models. This

is unsurprising, the simple enrichment metric used along side the linear regression doesn't account for interactions between the different predictors. Further development of this method would be to construct the models around interaction terms between every combination of the predictors or to move away from using linear regression.

Previously, Ouyang *et al* used principle component analysis on 12 transcription factors to explain 70% of the variance of micro-array expression data.[74] Cheng and Gerstein have similarly used support vector regression on a set of 12 transcription factors and 7 histone modifications to explain 72% of the variance in their expression data.[105] The enrichment metric-based method developed here is an improvement on these previous methods; the standard linear regression explaining 78% of the variance of the RNA polymerase II binding data and the LASSO regression explaining $\sim$80% of the variance. While the effectiveness of the methods vary, the statistical redundancy between the transcription factor and histone modification data that Cheng and Gerstein highlight can be seen in the results of the methods developed here, especially the results of the LASSO regression on the human embryonic stem cell data.

The final chapter of this work was working on the difficult task of predicting hot-spots in protein-protein interactions. InterBasePro is a previously-developed database of protein-biomolecule interaction energies. This work focused on applying the information contained in InterBasePro to the prediction of hot-spot residues in protein-protein interactions. Firstly, InterBasePro was scaled using linear models to investigate the relationship it has to experimental alanine scanning mutagenesis data. This data was obtained first from the ASEdb[237] resource and then a new dataset was sourced directly from the literature. When compared to both datasets, InterBasePro performed poorly, with correlations of 0.29 and 0.22 when compared to the ASEdb and the alternative data respectively. A desolvation factor was then added to the InterBasePro data to try and account for a major effect on protein-protein interactions. The addition of this term decreased the correlation between the InterBasePro data and the experimental data to 0.0047. Compared to previous methods such as Foldef[216] and Robetta[220] the methods developed here perform poorly, despite Foldef, Robetta and InterBasePro all using computational descriptions of atomic interactions. While InterBasePro has a poor correlation with the experimental data, the correlation it does have means it could potentially be a base on which to develop further methods. While alone they would not be useful for predicting protein-protein interaction hot-spot residues, they do show slight potential and could provide a base upon which further methods could be based. A final investigation was made into using the number of close atomic contacts each residue in a protein-protein interface has with its partner proteins atoms. When compared to the experimental data, the Atom Contact data had

a correlation of 0.08. The correlation with the experimental data was vastly increased, however, when the experimental data was split into subsets based on the function of the proteins. Enzyme/Inhibitor complex data had a correlation of 0.25 with the Atom Contact data. Despite this poor correlation, and as with the results of using InterBasePro, the Atom Contact method could provide a base on which to develop further methods. A large problem in the development of methods to predict hot-spot residues is the lack of up-to-date sources of experimental alanine-scanning mutagenesis data. While ASEdb and BID[238] are the standards on which methods are tested, they lack much of the more recently released data. While progress has been made here into gathering some of this data from the literature, a large scale approach to this would be beneficial to the field.

Bibliography

[1] Lubert; Stryer. *Bioinformatics*. W. H. Freeman and Company, New York, 4th edition, 1997. ISBN 0-7167-2009-4.

[2] Harmit S Malik and Steven Henikoff. Phylogenomics of the nucleosome. *Nature structural biology*, 10(11):882–91, November 2003. ISSN 1072-8368. doi: 10.1038/nsb996. URL http://dx.doi.org/10.1038/nsb996.

[3] S Henikoff, K Ahmad, and H S Malik. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science (New York, N.Y.)*, 293(5532):1098–102, August 2001. ISSN 0036-8075. doi: 10.1126/science.1062939. URL http://www.ncbi.nlm.nih.gov/pubmed/11498581.

[4] David Gokhman, Ilana Livyatan, Badi Sri Sailaja, Shai Melcer, and Eran Meshorer. Multilayered chromatin analysis reveals E2f, Smad and Zfx as transcriptional regulators of histones. *Nature structural & molecular biology*, 20(1):119–26, January 2013. ISSN 1545-9985. doi: 10.1038/nsmb.2448. URL http://www.ncbi.nlm.nih.gov/pubmed/23222641.

[5] Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, February 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.02.005. URL http://www.ncbi.nlm.nih.gov/pubmed/17320507.

[6] Alejandro Vaquero, Michael B Scher, Dong Hoon Lee, Ann Sutton, Hwei-Ling Cheng, Frederick W Alt, Lourdes Serrano, Rolf Sternglanz, and Danny Reinberg. SirT2 is a histone deacetylase with preference for histone H4 Lys 16 during mitosis. *Genes & development*, 20(10):1256–61, May 2006. ISSN 0890-9369. doi: 10.1101/gad.1412706. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1472900\&tool=pmcentrez\&rendertype=abstract.

[7] N Li, Z Sun, and F Jiang. SOFTDOCK application to protein-protein interaction benchmark and CAPRI. *Proteins*, 69:801–808, 2007. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=PubMed\&dopt=Citation\&list_uids=17803216.

[8] Paul B Mason and Kevin Struhl. Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. *Molecular cell*, 17(6):831–40, March 2005. ISSN 1097-2765. doi: 10.1016/j.molcel.2005.02.017. URL http://www.ncbi.nlm.nih.gov/pubmed/15780939.

[9] Rushad Pavri, Bing Zhu, Guohong Li, Patrick Trojer, Subhrangsu Mandal, Ali Shilatifard, and Danny Reinberg. Histone H2B monoubiquitination functions cooperatively with FACT to regulate elongation by RNA polymerase II. *Cell*, 125(4):703–17, May 2006. ISSN 0092-8674. doi: 10.1016/j.cell.2006.04.029. URL http://www.ncbi.nlm.nih.gov/pubmed/16713563.

[10] Andrew J Bannister and Tony Kouzarides. Reversing histone methylation. *Nature*, 436 (7054):1103–6, August 2005. ISSN 1476-4687. doi: 10.1038/nature04048. URL http://dx.doi.org/10.1038/nature04048.

[11] Michael J Carrozza, Bing Li, Laurence Florens, Tamaki Suganuma, Selene K Swanson, Kenneth K Lee, Wei-Jong Shia, Scott Anderson, John Yates, Michael P Washburn, and Jerry L Workman. Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*, 123 (4):581–92, November 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2005.10.023. URL http://www.ncbi.nlm.nih.gov/pubmed/16286007.

[12] Michael-Christopher Keogh, Siavash K Kurdistani, Stephanie A Morris, Seong Hoon Ahn, Vladimir Podolny, Sean R Collins, Maya Schuldiner, Kayu Chin, Thanuja Punna, Natalie J Thompson, Charles Boone, Andrew Emili, Jonathan S Weissman, Timothy R Hughes, Brian D Strahl, Michael Grunstein, Jack F Greenblatt, Stephen Buratowski, and Nevan J Krogan. Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell*, 123(4):593–605, November 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2005.10.025. URL http://www.ncbi.nlm.nih.gov/pubmed/16286008.

[13] Eric Metzger, Melanie Wissmann, Na Yin, Judith M Müller, Robert Schneider, Antoine H F M Peters, Thomas Günther, Reinhard Buettner, and Roland Schüle. LSD1

demethylates repressive histone marks to promote androgen-receptor-dependent transcription. *Nature*, 437(7057):436–9, September 2005. ISSN 1476-4687. doi: 10.1038/nature04020. URL http://www.ncbi.nlm.nih.gov/pubmed/16079795.

[14] Moniaux Nicolas, Faivre Jamila, and Surinder Batra. *RNA Polymerase II Holoenzyme and Transcription Factors. In: Encyclopedia of Life Sciences (eLS)*. John Wiley & Sons, Ltd, Chichester, UK, May 2001. ISBN 0470016175. doi: 10.1002/9780470015902.a0003301.pub2.

[15] Stephen T Smale and James T Kadonaga. The RNA polymerase II core promoter. *Annual review of biochemistry*, 72:449–79, January 2003. ISSN 0066-4154. doi: 10.1146/annurev.biochem.72.121801.161520. URL http://www.ncbi.nlm.nih.gov/pubmed/12651739.

[16] Dirk Kostrewa, Mirijam E Zeller, Karim-Jean Armache, Martin Seizl, Kristin Leike, Michael Thomm, and Patrick Cramer. RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature*, 462(7271):323–30, November 2009. ISSN 1476-4687. doi: 10.1038/nature08548. URL http://www.ncbi.nlm.nih.gov/pubmed/19820686.

[17] Xin Liu, David A Bushnell, Dong Wang, Guillermo Calero, and Roger D Kornberg. Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism. *Science (New York, N.Y.)*, 327(5962):206–9, January 2010. ISSN 1095-9203. doi: 10.1126/science.1182015. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2813267\&tool=pmcentrez\&rendertype=abstract.

[18] Y Li, P M Flanagan, H Tschochner, and R D Kornberg. RNA polymerase II initiation factor interactions and transcription start site selection. *Science (New York, N.Y.)*, 263 (5148):805–7, February 1994. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/8303296.

[19] T S Pardee, C S Bangur, and A S Ponticelli. The N-terminal region of yeast TFIIB contains two adjacent functional domains involved in stable RNA polymerase II binding and transcription start site selection. *The Journal of biological chemistry*, 273(28):17859–64, July 1998. ISSN 0021-9258. URL http://www.ncbi.nlm.nih.gov/pubmed/9651390.

[20] G Orphanides, T Lagrange, and D Reinberg. The general transcription factors of RNA polymerase II. *Genes & development*, 10(21):2657–83, November 1996. ISSN 0890-9369. URL http://www.ncbi.nlm.nih.gov/pubmed/8946909.

[21] Q Yan, R J Moreland, J W Conaway, and R C Conaway. Dual roles for transcription factor IIF in promoter escape by RNA polymerase II. *The Journal of biological chemistry*, 274 (50):35668–75, December 1999. ISSN 0021-9258. URL http://www.ncbi.nlm.nih.gov/pubmed/10585446.

[22] Roger D Kornberg. Mediator and the mechanism of transcriptional activation. *Trends in biochemical sciences*, 30(5):235–9, May 2005. ISSN 0968-0004. doi: 10.1016/j.tibs.2005.03.011. URL http://www.ncbi.nlm.nih.gov/pubmed/15896740.

[23] T K Kim, R H Ebright, and D Reinberg. Mechanism of ATP-dependent promoter melting by transcription factor IIH. *Science (New York, N.Y.)*, 288(5470):1418–22, May 2000. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/10827951.

[24] Michael-Christopher Keogh, Eun-Jung Cho, Vladimir Podolny, and Stephen Buratowski. Kin28 is found within TFIIH and a Kin28-Ccl1-Tfb3 trimer complex with differential sensitivities to T-loop phosphorylation. *Molecular and cellular biology*, 22(5):1288–97, March 2002. ISSN 0270-7306. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=134711\&tool=pmcentrez\&rendertype=abstract.

[25] J Q Svejstrup, W J Feaver, and R D Kornberg. Subunits of yeast RNA polymerase II transcription factor TFIIH encoded by the CCL1 gene. *The Journal of biological chemistry*, 271(2):643–5, January 1996. ISSN 0021-9258. URL http://www.ncbi.nlm.nih.gov/pubmed/8557668.

[26] Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A M Semple, Martin S Taylor, Pär G Engström, Martin C Frith, Alistair R R Forrest, Wynand B Alkema, Sin Lam Tan, Charles Plessy, Rimantas Kodzius, Timothy Ravasi, Takeya Kasukawa, Shiro Fukuda, Mutsumi Kanamori-Katayama, Yayoi Kitazume, Hideya Kawaji, Chikatoshi Kai, Mari Nakamura, Hideaki Konno, Kenji Nakano, Salim Mottagui-Tabar, Peter Arner, Alessandra Chesi, Stefano Gustincich, Francesca Persichetti, Harukazu Suzuki, Sean M Grimmond, Christine A Wells, Valerio Orlando, Claes Wahlestedt, Edison T Liu, Matthias Harbers, Jun Kawai, Vladimir B Bajic, David A Hume, and Yoshihide Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics*, 38(6):626–35, June 2006. ISSN 1061-4036. doi: 10.1038/ng1789. URL http://www.ncbi.nlm.nih.gov/pubmed/16645617.

[27] Chuhu Yang, Eugene Bolotin, Tao Jiang, Frances M Sladek, and Ernest Martinez. Prevalence of the initiator over the TATA box in human and yeast

genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, 389(1):52–65, March 2007. ISSN 0378-1119. doi: 10.1016/j.gene.2006.09.029. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1955227\&tool=pmcentrez\&rendertype=abstract.

[28] Neeman Mohibullah and Steven Hahn. Site-specific cross-linking of TBP in vivo and in vitro reveals a direct functional interaction with the SAGA subunit Spt3. *Genes & development*, 22(21):2994–3006, November 2008. ISSN 0890-9369. doi: 10.1101/gad.1724408. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2577793\&tool=pmcentrez\&rendertype=abstract.

[29] A M Dudley, C Rougeulle, and F Winston. The Spt components of SAGA facilitate TBP binding to a promoter at a post-activator-binding step in vivo. *Genes & development*, 13(22):2940–5, November 1999. ISSN 0890-9369. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=317152\&tool=pmcentrez\&rendertype=abstract.

[30] Sukesh R Bhaumik and Michael R Green. Differential requirement of SAGA components for recruitment of TATA-box-binding protein to promoters in vivo. *Molecular and cellular biology*, 22(21):7365–71, November 2002. ISSN 0270-7306. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=135674\&tool=pmcentrez\&rendertype=abstract.

[31] Ho Sung Rhee and B Franklin Pugh. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 483(7389):295–301, March 2012. ISSN 1476-4687. doi: 10.1038/nature10799. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3306527\&tool=pmcentrez\&rendertype=abstract.

[32] Seth R Goldman, Richard H Ebright, and Bryce E Nickels. Direct detection of abortive RNA transcripts in vivo. *Science (New York, N.Y.)*, 324(5929):927–8, May 2009. ISSN 1095-9203. doi: 10.1126/science.1169237. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2718712\&tool=pmcentrez\&rendertype=abstract.

[33] M E Maxon, J A Goodrich, and R Tjian. Transcription factor IIE binds preferentially to RNA polymerase IIa and recruits TFIIH: a model for promoter clearance. *Genes & development*, 8(5):515–24, March 1994. ISSN 0890-9369. URL http://www.ncbi.nlm.nih.gov/pubmed/7926747.

[34] Vladimir Svetlov and Evgeny Nudler. Basic mechanism of transcription by RNA polymerase II. *Biochimica et biophysica acta*, 1829(1):20–8, January 2013. ISSN 0006-3002. doi: 10.1016/j.bbagrm.2012.08.009. URL http://www.ncbi.nlm.nih.gov/pubmed/22982365.

[35] Steven West, Natalia Gromak, and Nick J Proudfoot. Human 5' –> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature*, 432(7016):522–5, November 2004. ISSN 1476-4687. doi: 10.1038/nature03035. URL http://www.ncbi.nlm.nih.gov/pubmed/15565158.

[36] William F Marzluff, Eric J Wagner, and Robert J Duronio. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nature reviews. Genetics*, 9(11):843–54, November 2008. ISSN 1471-0064. doi: 10.1038/nrg2438. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2715827\&tool=pmcentrez\&rendertype=abstract.

[37] Patricia Richard and James L Manley. Transcription termination by nuclear RNA polymerases. *Genes & development*, 23(11):1247–69, June 2009. ISSN 1549-5477. doi: 10.1101/gad.1792809. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2763537\&tool=pmcentrez\&rendertype=abstract.

[38] Stefania Millevoi and Stéphan Vagner. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic acids research*, 38(9):2757–74, May 2010. ISSN 1362-4962. doi: 10.1093/nar/gkp1176. URL http://nar.oxfordjournals.org/content/38/9/2757.

[39] Jason N Kuehner, Erika L Pearson, and Claire Moore. Unravelling the means to an end: RNA polymerase II transcription termination. *Nature reviews. Molecular cell biology*, 12 (5):283–94, May 2011. ISSN 1471-0080. doi: 10.1038/nrm3098. URL http://www.ncbi.nlm.nih.gov/pubmed/21487437.

[40] Hannah E Mischo and Nick J Proudfoot. Disengaging polymerase: Terminating RNA polymerase II transcription in budding yeast. *Biochimica et biophysica acta*, 1829(1): 174–85, January 2013. ISSN 0006-3002. doi: 10.1016/j.bbagrm.2012.10.003. URL http://www.ncbi.nlm.nih.gov/pubmed/23085255.

[41] M Merika, A J Williams, G Chen, T Collins, and D Thanos. Recruitment of CBP/p300 by the IFN beta enhanceosome is required for synergistic activation of transcription.

Molecular cell, 1(2):277–87, January 1998. ISSN 1097-2765. URL http://www.ncbi.nlm.nih.gov/pubmed/9659924.

[42] Ty C Voss, R Louis Schiltz, Myong-Hee Sung, Paul M Yen, John A Stamatoyannopoulos, Simon C Biddie, Thomas A Johnson, Tina B Miranda, Sam John, and Gordon L Hager. Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. *Cell*, 146(4):544–54, August 2011. ISSN 1097-4172. doi: 10.1016/j.cell.2011.07.006. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3210475\&tool=pmcentrez\&rendertype=abstract.

[43] Daniel Panne, Tom Maniatis, and Stephen C Harrison. An atomic model of the interferon-beta enhanceosome. *Cell*, 129(6):1111–23, June 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.05.019. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2020837\&tool=pmcentrez\&rendertype=abstract.

[44] J V Falvo, D Thanos, and T Maniatis. Reversal of intrinsic DNA bends in the IFN beta gene enhancer by transcription factors and the architectural protein HMG I(Y). *Cell*, 83 (7):1101–11, December 1995. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/8548798.

[45] Serena Ghisletti, Iros Barozzi, Flore Mietton, Sara Polletti, Francesca De Santa, Elisa Venturini, Lorna Gregory, Lorne Lonie, Adeline Chew, Chia-Lin Wei, Jiannis Ragoussis, and Gioacchino Natoli. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity*, 32(3): 317–28, March 2010. ISSN 1097-4180. doi: 10.1016/j.immuni.2010.02.008. URL http://www.ncbi.nlm.nih.gov/pubmed/20206554.

[46] Rasmus Siersbæ k, Ronni Nielsen, Sam John, Myong-Hee Sung, Songjoon Baek, Anne Loft, Gordon L Hager, and Susanne Mandrup. Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis. *The EMBO journal*, 30(8):1459–72, April 2011. ISSN 1460-2075. doi: 10.1038/emboj.2011.65. URL http://dx.doi.org/10.1038/emboj.2011.65.

[47] D Thanos and T Maniatis. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell*, 83(7):1091–100, December 1995. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/8548797.

[48] M Merika and D Thanos. Enhanceosomes. *Current opinion in genetics & develop-*

ment, 11(2):205–8, April 2001. ISSN 0959-437X. URL http://www.ncbi.nlm.nih.gov/pubmed/11250145.

[49] Christina I Swanson, Nicole C Evans, and Scott Barolo. Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Developmental cell*, 18(3):359–70, March 2010. ISSN 1878-1551. doi: 10.1016/j.devcel.2009.12.026. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2847355\&tool=pmcentrez\&rendertype=abstract.

[50] Nicola Reynolds, Paulina Latos, Antony Hynes-Allen, Remco Loos, Donna Leaford, Aoife O'Shaughnessy, Olukunbi Mosaku, Jason Signolet, Philip Brennecke, Tüzer Kalkan, Ita Costello, Peter Humphreys, William Mansfield, Kentaro Nakagawa, John Strouboulis, Axel Behrens, Paul Bertone, and Brian Hendrich. NuRD suppresses pluripotency gene expression to promote transcriptional heterogeneity and lineage commitment. *Cell stem cell*, 10(5):583–94, May 2012. ISSN 1875-9777. doi: 10.1016/j.stem.2012.02.020. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3402183\&tool=pmcentrez\&rendertype=abstract.

[51] Hitoshi Niwa. How is pluripotency determined and maintained? *Development (Cambridge, England)*, 134(4):635–46, February 2007. ISSN 0950-1991. doi: 10.1242/dev.02787. URL http://dev.biologists.org/content/134/4/635.short.

[52] H Niwa, J Miyazaki, and A G Smith. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nature genetics*, 24(4):372–6, April 2000. ISSN 1061-4036. doi: 10.1038/74199. URL http://www.ncbi.nlm.nih.gov/pubmed/10742100.

[53] Kai Wang, Satyaki Sengupta, Luca Magnani, Catherine A Wilson, R William Henry, and Jason G Knott. Brg1 is required for Cdx2-mediated repression of Oct4 expression in mouse blastocysts. *PloS one*, 5(5):e10622, January 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0010622. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2868905\&tool=pmcentrez\&rendertype=abstract.

[54] Bradley E Bernstein, Tarjei S Mikkelsen, Xiaohui Xie, Michael Kamal, Dana J Huebert, James Cuff, Ben Fry, Alex Meissner, Marius Wernig, Kathrin Plath, Rudolf Jaenisch, Alexandre Wagschal, Robert Feil, Stuart L Schreiber, and Eric S Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125

(2):315–26, April 2006. ISSN 0092-8674. doi: 10.1016/j.cell.2006.02.041. URL http://www.ncbi.nlm.nih.gov/pubmed/16630819.

[55] Véronique Azuara, Pascale Perry, Stephan Sauer, Mikhail Spivakov, Helle F Jø rgensen, Rosalind M John, Mina Gouti, Miguel Casanova, Gary Warnes, Matthias Merkenschlager, and Amanda G Fisher. Chromatin signatures of pluripotent cell lines. *Nature cell biology*, 8(5):532–8, May 2006. ISSN 1465-7392. doi: 10.1038/ncb1403. URL http://www.ncbi.nlm.nih.gov/pubmed/16570078.

[56] Z Nawaz, C Baniahmad, T P Burris, D J Stillman, B W O'Malley, and M J Tsai. The yeast SIN3 gene product negatively regulates the activity of the human progesterone receptor and positively regulates the activities of GAL4 and the HAP1 activator. *Molecular & general genetics : MGG*, 245(6):724–33, December 1994. ISSN 0026-8925. URL http://www.ncbi.nlm.nih.gov/pubmed/7830720.

[57] M Vidal, R Strich, R E Esposito, and R F Gaber. RPD1 (SIN3/UME4) is required for maximal activation and repression of diverse yeast genes. *Molecular and cellular biology*, 11(12):6306–16, December 1991. ISSN 0270-7306. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=361824\&tool=pmcentrez\&rendertype=abstract.

[58] M Vidal and R F Gaber. RPD3 encodes a second factor required to achieve maximum positive and negative transcriptional states in Saccharomyces cerevisiae. *Molecular and cellular biology*, 11(12):6317–27, December 1991. ISSN 0270-7306. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=361826\&tool=pmcentrez\&rendertype=abstract.

[59] Jan Christian Bryne, Eivind Valen, Man-Hung Eric Tang, Troels Marstrand, Ole Winther, Isabelle da Piedade, Anders Krogh, Boris Lenhard, and Albin Sandelin. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research*, 36(Database issue):D102–6, January 2008. ISSN 1362-4962. doi: 10.1093/nar/gkm955. URL http://nar.oxfordjournals.org/cgi/content/abstract/36/suppl_1/D102.

[60] V Matys, E Fricke, R Geffers, E Gössling, M Haubrock, R Hehl, K Hornischer, D Karas, A E Kel, O V Kel-Margoulis, D-U Kloos, S Land, B Lewicki-Potapov, H Michael, R Münch, I Reuter, S Rotert, H Saxel, M Scheer, S Thiele, and E Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, 31(1):

374–8, January 2003. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=165555\&tool=pmcentrez\&rendertype=abstract.

[61] Janne Korhonen, Petri Martinmäki, Cinzia Pizzi, Pasi Rastas, and Esko Ukkonen. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics (Oxford, England)*, 25(23):3181–2, December 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp554. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2778336\&tool=pmcentrez\&rendertype=abstracthttp://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/23/3181.

[62] Andrija Tomovic and Edward J Oakeley. Position dependencies in transcription factor binding sites. *Bioinformatics (Oxford, England)*, 23(8):933–41, May 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm055. URL http://www.ncbi.nlm.nih.gov/pubmed/17308339.

[63] F Sanger and A R Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–8, May 1975. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/1100841.

[64] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Chun He Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-Bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari A Vogt, Greg A Volkmer, Shally H Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80, September 2005. ISSN 1476-4687. doi: 10.1038/nature03959. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1464427\&tool=pmcentrez\&rendertype=abstract.

[65] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith,

John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, Jonathan M Boutell, Jason Bryant, Richard J Carter, R Keira Cheetham, Anthony J Cox, Darren J Ellis, Michael R Flatbush, Niall A Gormley, Sean J Humphray, Leslie J Irving, Mirian S Karbelashvili, Scott M Kirk, Heng Li, Xiaohai Liu, Klaus S Maisinger, Lisa J Murray, Bojan Obradovic, Tobias Ost, Michael L Parkinson, Mark R Pratt, Isabelle M J Rasolonjatovo, Mark T Reed, Roberto Rigatti, Chiara Rodighiero, Mark T Ross, Andrea Sabot, Subramanian V Sankar, Aylwyn Scally, Gary P Schroth, Mark E Smith, Vincent P Smith, Anastassia Spiridou, Peta E Torrance, Svilen S Tzonev, Eric H Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D Alam, Carole Anastasi, Ify C Aniebo, David M D Bailey, Iain R Bancarz, Saibal Banerjee, Selena G Barbour, Primo A Baybayan, Vincent A Benoit, Kevin F Benson, Claire Bevis, Phillip J Black, Asha Boodhun, Joe S Brennan, John A Bridgham, Rob C Brown, Andrew A Brown, Dale H Buermann, Abass A Bundu, James C Burrows, Nigel P Carter, Nestor Castillo, Maria Chiara E Catenazzi, Simon Chang, R Neil Cooley, Natasha R Crake, Olubunmi O Dada, Konstantinos D Diakoumakos, Belen Dominguez-Fernandez, David J Earnshaw, Ugonna C Egbujor, David W Elmore, Sergey S Etchin, Mark R Ewan, Milan Fedurco, Louise J Fraser, Karin V Fuentes Fajardo, W Scott Furey, David George, Kimberley J Gietzen, Colin P Goddard, George S Golda, Philip A Granieri, David E Green, David L Gustafson, Nancy F Hansen, Kevin Harnish, Christian D Haudenschild, Narinder I Heyer, Matthew M Hims, Johnny T Ho, Adrian M Horgan, Katya Hoschler, Steve Hurwitz, Denis V Ivanov, Maria Q Johnson, Terena James, T A Huw Jones, Gyoung-Dong Kang, Tzvetana H Kerelska, Alan D Kersey, Irina Khrebtukova, Alex P Kindwall, Zoya Kingsbury, Paula I Kokko-Gonzales, Anil Kumar, Marc A Laurent, Cynthia T Lawley, Sarah E Lee, Xavier Lee, Arnold K Liao, Jennifer A Loch, Mitch Lok, Shujun Luo, Radhika M Mammen, John W Martin, Patrick G McCauley, Paul McNitt, Parul Mehta, Keith W Moon, Joe W Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M Novo, Michael J O'Neill, Mark A Osborne, Andrew Osnowski, Omead Ostadan, Lambros L Paraschos, Lea Pickering, Andrew C Pike, Alger C Pike, D Chris Pinkard, Daniel P Pliskin, Joe Podhasky, Victor J Quijano, Come Raczy, Vicki H Rae, Stephen R Rawlings, Ana Chiva Rodriguez, Phyllida M Roe, John Rogers, Maria C Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K Roth, Natalie J Rourke, Silke T Ruediger, Eli Rusman, Raquel M Sanches-Kuiper, Martin R Schenker, Josefina M Seoane, Richard J Shaw, Mitch K Shiver, Steven W Short, Ning L Sizto, Johannes P Sluis, Melanie A Smith, Jean Ernest Sohna Sohna, Eric J Spence, Kim Stevens, Neil Sutton, Lukasz

Szajkowski, Carolyn L Tregidgo, Gerardo Turcatti, Stephanie Vandevondele, Yuli Verhovsky, Selene M Virk, Suzanne Wakelin, Gregory C Walcott, Jingwen Wang, Graham J Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C Mullikin, Matthew E Hurles, Nick J McCooke, John S West, Frank L Oaks, Peter L Lundberg, David Klenerman, Richard Durbin, and Anthony J Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9, November 2008. ISSN 1476-4687. doi: 10.1038/nature07517. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2581791\&tool=pmcentrez\&rendertype=abstract.

[66] Kevin Judd McKernan, Heather E Peckham, Gina L Costa, Stephen F McLaughlin, Yutao Fu, Eric F Tsung, Christopher R Clouser, Cisyla Duncan, Jeffrey K Ichikawa, Clarence C Lee, Zheng Zhang, Swati S Ranade, Eileen T Dimalanta, Fiona C Hyland, Tanya D Sokolsky, Lei Zhang, Andrew Sheridan, Haoning Fu, Cynthia L Hendrickson, Bin Li, Lev Kotler, Jeremy R Stuart, Joel A Malek, Jonathan M Manning, Alena A Antipova, Damon S Perez, Michael P Moore, Kathleen C Hayashibara, Michael R Lyons, Robert E Beaudoin, Brittany E Coleman, Michael W Laptewicz, Adam E Sannicandro, Michael D Rhodes, Rajesh K Gottimukkala, Shan Yang, Vineet Bafna, Ali Bashir, Andrew MacBride, Can Alkan, Jeffrey M Kidd, Evan E Eichler, Martin G Reese, Francisco M De La Vega, and Alan P Blanchard. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research*, 19(9):1527–41, September 2009. ISSN 1549-5469. doi: 10.1101/gr.091868.109. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2752135\&tool=pmcentrez\&rendertype=abstract.

[67] Philippe Collas. The current state of chromatin immunoprecipitation. *Molecular biotechnology*, 45(1):87–100, May 2010. ISSN 1559-0305. doi: 10.1007/s12033-009-9239-8. URL http://www.ncbi.nlm.nih.gov/pubmed/20077036.

[68] Alan P Boyle, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–22, January 2008. ISSN 1097-4172. doi: 10.1016/j.cell.2007.12.014. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2669738\&tool=pmcentrez\&rendertype=abstract.

[69] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–

60, July 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp324. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234\&tool=pmcentrez\&rendertype=abstract.

[70] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10 (3):R25, January 2009. ISSN 1465-6914. doi: 10.1186/gb-2009-10-3-r25. URL http://genomebiology.com/2009/10/3/R25.

[71] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nussbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9):R137, January 2008. ISSN 1465-6914. doi: 10.1186/gb-2008-9-9-r137. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2592715\&tool=pmcentrez\&rendertype=abstract.

[72] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology*, 27(1):66–75, January 2009. ISSN 1546-1696. doi: 10.1038/nbt.1518. URL http://www.ncbi.nlm.nih.gov/pubmed/19122651.

[73] Xin Feng, Robert Grossman, and Lincoln Stein. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC bioinformatics*, 12:139, January 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-139. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3103446\&tool=pmcentrez\&rendertype=abstract.

[74] Zhengqing Ouyang, Qing Zhou, and Wing Hung Wong. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21521–6, December 2009. ISSN 1091-6490. doi: 10.1073/pnas.0904863106. URL http://www.pnas.org/cgi/content/abstract/106/51/21521.

[75] Chao Cheng and Mark Gerstein. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic acids research*, 40(2):553–68, January 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr752. URL http://nar.oxfordjournals.org/content/40/2/553http:

//www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3258143\&tool=pmcentrez\&rendertype=abstract.

[76] Chao Cheng, Roger Alexander, Renqiang Min, Jing Leng, Kevin Y Yip, Joel Rozowsky, Koon-Kiu Yan, Xianjun Dong, Sarah Djebali, Yijun Ruan, Carrie a Davis, Piero Carninci, Timo Lassman, Thomas R Gingeras, Roderic Guigó, Ewan Birney, Zhiping Weng, Michael Snyder, and Mark Gerstein. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome research*, 22(9):1658–67, September 2012. ISSN 1549-5469. doi: 10.1101/gr.136838.111. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3431483\&tool=pmcentrez\&rendertype=abstract.

[77] David M Mosser and Justin P Edwards. Exploring the full spectrum of macrophage activation. *Nature reviews. Immunology*, 8(12):958–69, December 2008. ISSN 1474-1741. doi: 10.1038/nri2448. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2724991\&tool=pmcentrez\&rendertype=abstract.

[78] B Passlick, D Flieger, and H W Ziegler-Heitbrock. Identification and characterization of a novel monocyte subpopulation in human peripheral blood. *Blood*, 74(7):2527–34, November 1989. ISSN 0006-4971. URL http://www.ncbi.nlm.nih.gov/pubmed/2478233.

[79] G B Mackaness. Cellular immunity and the parasite. *Advances in experimental medicine and biology*, 93:65–73, January 1977. ISSN 0065-2598. URL http://www.ncbi.nlm.nih.gov/pubmed/339685.

[80] John J O'Shea and Peter J Murray. Cytokine signaling modules in inflammatory responses. *Immunity*, 28(4):477–87, April 2008. ISSN 1097-4180. doi: 10.1016/j.immuni.2008.03.002. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2782488\&tool=pmcentrez\&rendertype=abstract.

[81] David C Dale, Laurence Boxer, and W Conrad Liles. The phagocytes: neutrophils and monocytes. *Blood*, 112(4):935–45, August 2008. ISSN 1528-0020. doi: 10.1182/blood-2007-12-077917. URL http://www.ncbi.nlm.nih.gov/pubmed/18684880.

[82] David M Mosser. The many faces of macrophage activation. *Journal of leukocyte biology*, 73(2):209–12, February 2003. ISSN 0741-5400. URL http://www.ncbi.nlm.nih.gov/pubmed/12554797.

[83] Mark Lucas, Xia Zhang, Vikram Prasanna, and David M Mosser. ERK activation following macrophage FcgammaR ligation leads to chromatin modifications at the IL-10 locus. *Journal of immunology (Baltimore, Md. : 1950)*, 175(1):469–77, July 2005. ISSN 0022-1767. URL http://www.ncbi.nlm.nih.gov/pubmed/15972681.

[84] Justin P Edwards, Xia Zhang, Kenneth A Frauwirth, and David M Mosser. Biochemical and functional characterization of three activated macrophage populations. *Journal of leukocyte biology*, 80(6):1298–307, December 2006. ISSN 0741-5400. doi: 10.1189/jlb.0406249. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2642590\&tool=pmcentrez\&rendertype=abstract.

[85] J S Gerber and D M Mosser. Reversing lipopolysaccharide toxicity by ligating the macrophage Fc gamma receptors. *Journal of immunology (Baltimore, Md. : 1950)*, 166(11):6861–8, June 2001. ISSN 0022-1767. URL http://www.ncbi.nlm.nih.gov/pubmed/11359846.

[86] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, 38(4):576–89, May 2010. ISSN 1097-4164. doi: 10.1016/j.molcel.2010.05.004. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2898526\&tool=pmcentrez\&rendertype=abstract.

[87] Gioacchino Natoli. Maintaining cell identity through global control of genomic organization. *Immunity*, 33(1):12–24, July 2010. ISSN 1097-4180. doi: 10.1016/j.immuni.2010.07.006. URL http://www.ncbi.nlm.nih.gov/pubmed/20643336.

[88] Gioacchino Natoli. Control of NF-kappaB-dependent transcriptional responses by chromatin organization. *Cold Spring Harbor perspectives in biology*, 1 (4):a000224, October 2009. ISSN 1943-0264. doi: 10.1101/cshperspect.a000224. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2773620\&tool=pmcentrez\&rendertype=abstract.

[89] Toby Lawrence and Gioacchino Natoli. Transcriptional regulation of macrophage polarization: enabling diversity with identity. *Nature reviews. Immunology*, 11(11):750–61, November 2011. ISSN 1474-1741. doi: 10.1038/nri3088. URL http://www.ncbi.nlm.nih.gov/pubmed/22025054.

[90] Katia Basso and Riccardo Dalla-Favera. BCL6: master regulator of the germinal center reaction and key oncogene in B cell lymphomagenesis. *Advances in immunology*, 105: 193–210, January 2010. ISSN 1557-8445. doi: 10.1016/S0065-2776(10)05007-8. URL http://www.ncbi.nlm.nih.gov/pubmed/20510734.

[91] Timothy Ravasi, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B Bajic, Kai Tan, Altuna Akalin, Sebastian Schmeier, Mutsumi Kanamori-Katayama, Nicolas Bertin, Piero Carninci, Carsten O Daub, Alistair R R Forrest, Julian Gough, Sean Grimmond, Jung-Hoon Han, Takehiro Hashimoto, Winston Hide, Oliver Hofmann, Atanas Kamburov, Mandeep Kaur, Hideya Kawaji, Atsutaka Kubosaki, Timo Lassmann, Erik van Nimwegen, Cameron Ross MacPherson, Chihiro Ogawa, Aleksandar Radovanovic, Ariel Schwartz, Rohan D Teasdale, Jesper Tegnér, Boris Lenhard, Sarah A Teichmann, Takahiro Arakawa, Noriko Ninomiya, Kayoko Murakami, Michihira Tagami, Shiro Fukuda, Kengo Imamura, Chikatoshi Kai, Ryoko Ishihara, Yayoi Kitazume, Jun Kawai, David A Hume, Trey Ideker, and Yoshihide Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–52, March 2010. ISSN 1097-4172. doi: 10.1016/j.cell.2010.01.044. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2836267\&tool=pmcentrez\&rendertype=abstract.

[92] Grant D Barish, Ruth T Yu, Malith Karunasiri, Corinne B Ocampo, Jesse Dixon, Chris Benner, Alexander L Dent, Rajendra K Tangirala, and Ronald M Evans. Bcl-6 and NF-kappaB cistromes mediate opposing regulation of the innate immune response. *Genes & development*, 24(24):2760–5, December 2010. ISSN 1549-5477. doi: 10.1101/gad.1998010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3003193\&tool=pmcentrez\&rendertype=abstract.

[93] Oksana R Bereshchenko, Wei Gu, and Riccardo Dalla-Favera. Acetylation inactivates the transcriptional repressor BCL6. *Nature genetics*, 32(4):606–13, December 2002. ISSN 1061-4036. doi: 10.1038/ng1018. URL http://www.ncbi.nlm.nih.gov/pubmed/12402037.

[94] Gioacchino Natoli, Serena Ghisletti, and Iros Barozzi. The genomic landscapes of inflammation. *Genes & development*, 25(2):101–6, January 2011. ISSN 1549-5477. doi: 10.1101/gad.2018811. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3022255\&tool=pmcentrez\&rendertype=abstract.

[95] Christopher K Glass and Kaoru Saijo. Nuclear receptor transrepression pathways that regulate inflammation in macrophages and T cells. *Nature reviews. Immunology*, 10(5): 365–76, May 2010. ISSN 1474-1741. doi: 10.1038/nri2748. URL http://www.ncbi.nlm.nih.gov/pubmed/20414208.

[96] E R Stanley. Action of the colony-stimulating factor, CSF-1. *Ciba Foundation symposium*, 118:29–41, January 1986. ISSN 0300-5208. URL http://www.ncbi.nlm.nih.gov/pubmed/3015514.

[97] Tanawan Kummalue and Alan D Friedman. Cross-talk between regulators of myeloid development: C/EBPalpha binds and activates the promoter of the PU.1 gene. *Journal of leukocyte biology*, 74(3):464–70, September 2003. ISSN 0741-5400. URL http://www.ncbi.nlm.nih.gov/pubmed/12949251.

[98] Christine Yeamans, Dehua Wang, Ido Paz-Priel, Bruce E Torbett, Daniel G Tenen, and Alan D Friedman. C/EBPalpha binds and activates the PU.1 distal enhancer to induce monocyte lineage commitment. *Blood*, 110(9):3136–42, November 2007. ISSN 0006-4971. doi: 10.1182/blood-2007-03-080291. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2200910\&tool=pmcentrez\&rendertype=abstract.

[99] R J Christy, K H Kaestner, D E Geiman, and M D Lane. CCAAT/enhancer binding protein gene promoter: binding of nuclear factors during differentiation of 3T3-L1 preadipocytes. *Proceedings of the National Academy of Sciences of the United States of America*, 88 (6):2593–7, March 1991. ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=51279\&tool=pmcentrez\&rendertype=abstract.

[100] Satoshi Iida, Rie Watanabe-Fukunaga, Shigekazu Nagata, and Rikiro Fukunaga. Essential role of C/EBPalpha in G-CSF-induced transcriptional activation and chromatin modification of myeloid-specific genes. *Genes to cells : devoted to molecular & cellular mechanisms*, 13(4):313–27, April 2008. ISSN 1365-2443. doi: 10.1111/j.1365-2443.2008.01173.x. URL http://www.ncbi.nlm.nih.gov/pubmed/18363963.

[101] Bo T Porse, David Bryder, Kim Theilgaard-Mönch, Marie S Hasemann, Kristina Anderson, Inge Damgaard, Sten Eirik W Jacobsen, and Claus Nerlov. Loss of C/EBP alpha cell cycle control increases myeloid progenitor proliferation and transforms the neutrophil granulocyte lineage. *The Journal of experimental medicine*, 202(1):85–96, July 2005. ISSN 0022-1007. doi: 10.1084/jem.

20050067. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2212897\&tool=pmcentrez\&rendertype=abstract.

[102] Daniela Ruffell, Foteini Mourkioti, Adriana Gambardella, Peggy Kirstetter, Rodolphe G Lopez, Nadia Rosenthal, and Claus Nerlov. A CREB-C/EBPbeta cascade induces M2 macrophage-specific gene expression and promotes muscle injury repair. *Proceedings of the National Academy of Sciences of the United States of America*, 106(41):17475–80, October 2009. ISSN 1091-6490. doi: 10.1073/pnas.0908641106. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2762675\&tool=pmcentrez\&rendertype=abstract.

[103] H M Hu, M Baer, S C Williams, P F Johnson, and R C Schwartz. Redundancy of C/EBP alpha, -beta, and -delta in supporting the lipopolysaccharide-induced transcription of IL-6 and monocyte chemoattractant protein-1. *Journal of immunology (Baltimore, Md. : 1950)*, 160(5):2334–42, March 1998. ISSN 0022-1767. URL http://www.ncbi.nlm.nih.gov/pubmed/9498774.

[104] Letetia C Jones, Meng-Liang Lin, Shih-Shun Chen, Utz Krug, Wolf-K Hofmann, Stephen Lee, Ying-Hue Lee, and H Phillip Koeffler. Expression of C/EBPbeta from the C/ebpalpha gene locus is sufficient for normal hematopoiesis in vivo. *Blood*, 99(6):2032–6, March 2002. ISSN 0006-4971. URL http://www.ncbi.nlm.nih.gov/pubmed/11877276.

[105] C. Cheng, R. Alexander, R. Min, J. Leng, K. Y. Yip, J. Rozowsky, K.-K. Yan, X. Dong, S. Djebali, Y. Ruan, C. A. Davis, P. Carninci, T. Lassman, T. R. Gingeras, R. Guigo, E. Birney, Z. Weng, M. Snyder, and M. Gerstein. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Research*, 22(9):1658–1667, September 2012. ISSN 1088-9051. doi: 10.1101/gr.136838.111. URL http://genome.cshlp.org/cgi/content/abstract/22/9/1658.

[106] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Rolf N Muertter, Michelle Holko, Oluwabukunmi Ayanbule, Andrey Yefanov, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic acids research*, 39(Database issue):D1005–10, January 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1184. URL http://nar.oxfordjournals.org/cgi/content/abstract/39/suppl_1/D1005.

[107] Mathias Leddin, Chiara Perrod, Maarten Hoogenkamp, Saeed Ghani, Salam Assi, Sven Heinz, Nicola K Wilson, George Follows, Jörg Schönheit, Lena Vockentanz, Ali M Mosammam, Wei Chen, Daniel G Tenen, David R Westhead, Berthold Göttgens, Constanze Bonifer, and Frank Rosenbauer. Two distinct auto-regulatory loops operate at the PU.1 locus in B cells and myeloid cells. *Blood*, 117(10):2827–38, March 2011. ISSN 1528-0020. doi: 10.1182/blood-2010-08-302976. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3062295\&tool=pmcentrez\&rendertype=abstract.

[108] Jacques Rougemont and Felix Naef. Computational analysis of protein-DNA interactions from ChIP-seq data. *Methods in molecular biology (Clifton, N.J.)*, 786:263–73, January 2012. ISSN 1940-6029. doi: 10.1007/978-1-61779-292-2_16. URL http://www.springerlink.com/content/w114553547548v30/.

[109] Elizabeth G. Wilbanks and Marc T. Facciotti. Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. *PLoS ONE*, 5(7):e11471, July 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0011471. URL http://dx.plos.org/10.1371/journal.pone.0011471.

[110] Kim D Pruitt, Tatiana Tatusova, William Klimke, and Donna R Maglott. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic acids research*, 37(Database issue):D32–6, January 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn721. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686572\&tool=pmcentrez\&rendertype=abstract.

[111] Morten Rye, På l Sæ trom, Tony Hå ndstad, and Finn Drablø s. Clustered ChIP-Seq-defined transcription factor binding sites and histone modifications map distinct classes of regulatory elements. *BMC biology*, 9(1):80, January 2011. ISSN 1741-7007. doi: 10.1186/1741-7007-9-80. URL http://www.biomedcentral.com/1741-7007/9/80.

[112] R Development Core Team. R: A Language and Environment for Statistical Computing, 2009. URL http://www.r-project.org.

[113] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9, May 2000. ISSN 1061-4036. doi: 10.

1038/75556. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037419\&tool=pmcentrez\&rendertype=abstract.

[114] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, January 2004. ISSN 1465-6914. doi: 10.1186/gb-2004-5-10-r80. URL http://genomebiology.com/2004/5/10/R80.

[115] Timothy Ravasi, Christine A Wells, and David A Hume. Systems biology of transcription control in macrophages. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 29(12):1215–26, December 2007. ISSN 0265-9247. doi: 10.1002/bies.20683. URL http://www.ncbi.nlm.nih.gov/pubmed/18008376.

[116] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50, October 2005. ISSN 0027-8424. doi: 10.1073/pnas.0506580102. URL http://www.pnas.org/cgi/content/abstract/102/43/15545.

[117] Chao Cheng, Renqiang Min, and Mark Gerstein. A Probabilistic Method for identifying Transcription Factor Target Genes from ChIP-Seq Binding Profiles. *Bioinformatics (Oxford, England)*, 27(23):3221–3227, October 2011. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr552. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/27/23/3221.

[118] Brian C Haynes, Ezekiel J Maier, Michael H Kramer, Patricia I Wang, Holly Brown, and Michael R Brent. Mapping Functional Transcription Factor Networks from Gene Expression Data. *Genome research*, May 2013. ISSN 1549-5469. doi: 10.1101/gr.150904.112. URL http://www.ncbi.nlm.nih.gov/pubmed/23636944.

[119] Francesco Romeo, Francesco Costanzo, and Massimiliano Agostini. Embryonic stem

cells and inducible pluripotent stem cells: two faces of the same coin? *Aging*, December 2012. ISSN 1945-4589. URL http://www.ncbi.nlm.nih.gov/pubmed/23248145.

[120] Nishanthi Gangadaran and Jannet Vennila James. Gene interaction studies in cellular reprogramming of adult stem cells to embyronic like stem cells. *Bioinformation*, 8(19):912–5, January 2012. ISSN 0973-2063. doi: 10.6026/97320630008912. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3488832\&tool=pmcentrez\&rendertype=abstract.

[121] Shinichi Matsumoto. Islet cell transplantation for Type 1 diabetes. *Journal of diabetes*, 2 (1):16–22, March 2010. ISSN 1753-0407. doi: 10.1111/j.1753-0407.2009.00048.x. URL http://www.ncbi.nlm.nih.gov/pubmed/20923470.

[122] Dali Yang, Zhi-Jian Zhang, Michael Oldenburg, Melvin Ayala, and Su-Chun Zhang. Human embryonic stem cell-derived dopaminergic neurons reverse functional deficit in parkinsonian rats. *Stem cells (Dayton, Ohio)*, 26(1):55–63, January 2008. ISSN 1549-4918. doi: 10.1634/stemcells.2007-0494. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2707927\&tool=pmcentrez\&rendertype=abstract.

[123] Shankar J Chinta and Julie K Andersen. Prospects and challenges for the use of stem cell technologies to develop novel therapies for Parkinson disease. *Cell cycle (Georgetown, Tex.)*, 10(24):4179–80, December 2011. ISSN 1551-4005. doi: 10.4161/cc.10.24.18835. URL http://www.ncbi.nlm.nih.gov/pubmed/22157187.

[124] Natasha Y Frank, Tobias Schatton, and Markus H Frank. The therapeutic promise of the cancer stem cell concept. *The Journal of clinical investigation*, 120(1):41–50, January 2010. ISSN 1558-8238. doi: 10.1172/JCI41004. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2798700\&tool=pmcentrez\&rendertype=abstract.

[125] Haruki Sekiguchi, Masaaki Ii, and Douglas W Losordo. The relative potency and safety of endothelial progenitor cells and unselected mononuclear cells for recovery from myocardial infarction and ischemia. *Journal of cellular physiology*, 219(2):235–42, May 2009. ISSN 1097-4652. doi: 10.1002/jcp.21672. URL http://www.ncbi.nlm.nih.gov/pubmed/19115244.

[126] L Crews, C Patrick, A Adame, E Rockenstein, and E Masliah. Modulation of aberrant CDK5 signaling rescues impaired neurogenesis in models of Alzheimer's disease. *Cell*

death & disease, 2:e120, January 2011. ISSN 2041-4889. doi: 10.1038/cddis.2011.2. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3101702\&tool=pmcentrez\&rendertype=abstract.

[127] Jia-Chi Yeo and Huck-Hui Ng. The transcriptional regulation of pluripotency. *Cell research*, 23(1):20–32, December 2012. ISSN 1748-7838. doi: 10.1038/cr.2012.172. URL http://dx.doi.org/10.1038/cr.2012.172.

[128] M H Rosner, M A Vigano, K Ozato, P M Timmons, F Poirier, P W Rigby, and L M Staudt. A POU-domain transcription factor in early stem cells and germ cells of the mammalian embryo. *Nature*, 345(6277):686–92, June 1990. ISSN 0028-0836. doi: 10.1038/345686a0. URL http://www.ncbi.nlm.nih.gov/pubmed/1972777.

[129] H R Schöler, G R Dressler, R Balling, H Rohdewohld, and P Gruss. Oct-4: a germline-specific transcription factor mapping to the mouse t-complex. *The EMBO journal*, 9 (7):2185–95, July 1990. ISSN 0261-4189. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=551941\&tool=pmcentrez\&rendertype=abstract.

[130] Ariel A Avilion, Silvia K Nicolis, Larysa H Pevny, Lidia Perez, Nigel Vivian, and Robin Lovell-Badge. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes & development*, 17(1):126–40, January 2003. ISSN 0890-9369. doi: 10.1101/gad.224503. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=195970\&tool=pmcentrez\&rendertype=abstract.

[131] H Yuan, N Corbi, C Basilico, and L Dailey. Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. *Genes & development*, 9(21):2635–45, November 1995. ISSN 0890-9369. URL http://www.ncbi.nlm.nih.gov/pubmed/7590241.

[132] M Nishimoto, A Fukushima, A Okuda, and M Muramatsu. The gene for the embryonic stem cell coactivator UTF1 carries a regulatory element which selectively interacts with a complex composed of Oct-3/4 and Sox-2. *Molecular and cellular biology*, 19(8): 5453–65, August 1999. ISSN 0270-7306. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=84387\&tool=pmcentrez\&rendertype=abstract.

[133] Yoshimi Tokuzawa, Eiko Kaiho, Masayoshi Maruyama, Kazutoshi Takahashi, Kaoru Mitsui, Mitsuyo Maeda, Hitoshi Niwa, and Shinya Yamanaka. Fbx15 is a novel target of Oct3/4 but is dispensable for embryonic stem cell self-renewal and mouse

development. *Molecular and cellular biology*, 23(8):2699–708, April 2003. ISSN 0270-7306. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=152544\&tool=pmcentrez\&rendertype=abstract.

[134] Yuhki Nakatake, Nobutaka Fukui, Yuko Iwamatsu, Shinji Masui, Kadue Takahashi, Rika Yagi, Kiyohito Yagi, Jun-Ichi Miyazaki, Ryo Matoba, Minoru S H Ko, and Hitoshi Niwa. Klf4 cooperates with Oct3/4 and Sox2 to activate the Lefty1 core promoter in embryonic stem cells. *Molecular and cellular biology*, 26(20):7772–82, October 2006. ISSN 0270-7306. doi: 10.1128/MCB.00468-06. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1636862\&tool=pmcentrez\&rendertype=abstract.

[135] Takao Kuroda, Masako Tada, Hiroshi Kubota, Hironobu Kimura, Shin-ya Hatano, Hirofumi Suemori, Norio Nakatsuji, and Takashi Tada. Octamer and Sox elements are required for transcriptional cis regulation of Nanog gene expression. *Molecular and cellular biology*, 25(6):2475–85, March 2005. ISSN 0270-7306. doi: 10.1128/MCB.25.6.2475-2485.2005. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1061601\&tool=pmcentrez\&rendertype=abstract.

[136] David J Rodda, Joon-Lin Chew, Leng-Hiong Lim, Yuin-Han Loh, Bei Wang, Huck-Hui Ng, and Paul Robson. Transcriptional regulation of nanog by OCT4 and SOX2. *The Journal of biological chemistry*, 280(26):24731–7, July 2005. ISSN 0021-9258. doi: 10.1074/jbc.M502573200. URL http://www.ncbi.nlm.nih.gov/pubmed/15860457.

[137] Sayaka Okumura-Nakanishi, Motoki Saito, Hitoshi Niwa, and Fuyuki Ishikawa. Oct-3/4 and Sox2 regulate Oct-3/4 gene in embryonic stem cells. *The Journal of biological chemistry*, 280(7):5307–17, February 2005. ISSN 0021-9258. doi: 10.1074/jbc.M410015200. URL http://www.ncbi.nlm.nih.gov/pubmed/15557334.

[138] Mizuho Tomioka, Masazumi Nishimoto, Satoru Miyagi, Tomoko Katayanagi, Nobutaka Fukui, Hitoshi Niwa, Masami Muramatsu, and Akihiko Okuda. Identification of Sox-2 regulatory region which is under the control of Oct-3/4-Sox-2 complex. *Nucleic acids research*, 30(14):3202–13, July 2002. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=135755\&tool=pmcentrez\&rendertype=abstract.

[139] Joon-Lin Chew, Yuin-Han Loh, Wensheng Zhang, Xi Chen, Wai-Leong Tam, Leng-Siew Yeap, Pin Li, Yen-Sin Ang, Bing Lim, Paul Robson, and Huck-Hui Ng. Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem

cells. *Molecular and cellular biology*, 25(14):6031–46, July 2005. ISSN 0270-7306. doi: 10.1128/MCB.25.14.6031-6046.2005. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1168830\&tool=pmcentrez\&rendertype=abstract.

[140] Ian Chambers, Jose Silva, Douglas Colby, Jennifer Nichols, Bianca Nijmeijer, Morag Robertson, Jan Vrana, Ken Jones, Lars Grotewold, and Austin Smith. Nanog safeguards pluripotency and mediates germline development. *Nature*, 450(7173):1230–4, December 2007. ISSN 1476-4687. doi: 10.1038/nature06403. URL http://www.ncbi.nlm.nih.gov/pubmed/18097409.

[141] Kaoru Mitsui, Yoshimi Tokuzawa, Hiroaki Itoh, Kohichi Segawa, Mirei Murakami, Kazutoshi Takahashi, Masayoshi Maruyama, Mitsuyo Maeda, and Shinya Yamanaka. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell*, 113(5):631–42, May 2003. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/12787504.

[142] Jose Silva, Jennifer Nichols, Thorold W Theunissen, Ge Guo, Anouk L van Oosten, Ornella Barrandon, Jason Wray, Shinya Yamanaka, Ian Chambers, and Austin Smith. Nanog is the gateway to the pluripotent ground state. *Cell*, 138(4):722–37, August 2009. ISSN 1097-4172. doi: 10.1016/j.cell.2009.07.039. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3437554\&tool=pmcentrez\&rendertype=abstract.

[143] Tongxiang Lin, Connie Chao, Shin'ichi Saito, Sharlyn J Mazur, Maureen E Murphy, Ettore Appella, and Yang Xu. p53 induces differentiation of mouse embryonic stem cells by suppressing Nanog expression. *Nature cell biology*, 7(2):165–71, February 2005. ISSN 1465-7392. doi: 10.1038/ncb1211. URL http://www.ncbi.nlm.nih.gov/pubmed/15619621.

[144] S Mora-Castilla, J R Tejedo, A Hmadcha, G M Cahuana, F Martín, B Soria, and F J Bedoya. Nitric oxide repression of Nanog promotes mouse embryonic stem cell differentiation. *Cell death and differentiation*, 17(6):1025–33, June 2010. ISSN 1476-5403. doi: 10.1038/cdd.2009.204. URL http://www.ncbi.nlm.nih.gov/pubmed/20075941.

[145] Yuin-Han Loh, Qiang Wu, Joon-Lin Chew, Vinsensius B Vega, Weiwei Zhang, Xi Chen, Guillaume Bourque, Joshy George, Bernard Leong, Jun Liu, Kee-Yew Wong, Ken W Sung, Charlie W H Lee, Xiao-Dong Zhao, Kuo-Ping Chiu, Leonard Lipovich, Vladimir A Kuznetsov, Paul Robson, Lawrence W Stanton, Chia-Lin Wei, Yijun Ruan, Bing Lim,

and Huck-Hui Ng. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature genetics*, 38(4):431–40, April 2006. ISSN 1061-4036. doi: 10.1038/ng1760. URL http://www.ncbi.nlm.nih.gov/pubmed/16518401.

[146] Laurie A Boyer, Tong Ihn Lee, Megan F Cole, Sarah E Johnstone, Stuart S Levine, Jacob P Zucker, Matthew G Guenther, Roshan M Kumar, Heather L Murray, Richard G Jenner, David K Gifford, Douglas A Melton, Rudolf Jaenisch, and Richard A Young. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–56, September 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2005.08.020. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3006442\&tool=pmcentrez\&rendertype=abstract.

[147] Jianlong Wang, Sridhar Rao, Jianlin Chu, Xiaohua Shen, Dana N Levasseur, Thorold W Theunissen, and Stuart H Orkin. A protein interaction network for pluripotency of embryonic stem cells. *Nature*, 444(7117):364–8, November 2006. ISSN 1476-4687. doi: 10.1038/nature05284. URL http://www.ncbi.nlm.nih.gov/pubmed/17093407.

[148] Jiancong Liang, Ma Wan, Yi Zhang, Peili Gu, Huawei Xin, Sung Yun Jung, Jun Qin, Jiemin Wong, Austin J Cooney, Dan Liu, and Zhou Songyang. Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. *Nature cell biology*, 10(6):731–9, June 2008. ISSN 1476-4679. doi: 10.1038/ncb1736. URL http://www.ncbi.nlm.nih.gov/pubmed/18454139.

[149] Qiang Wu, Xi Chen, Jinqiu Zhang, Yuin-Han Loh, Teck-Yew Low, Weiwei Zhang, Wensheng Zhang, Siu-Kwan Sze, Bing Lim, and Huck-Hui Ng. Sall4 interacts with Nanog and co-occupies Nanog genomic sites in embryonic stem cells. *The Journal of biological chemistry*, 281(34):24090–4, August 2006. ISSN 0021-9258. doi: 10.1074/jbc.C600122200. URL http://www.ncbi.nlm.nih.gov/pubmed/16840789.

[150] Debbie L C van den Berg, Tim Snoek, Nick P Mullin, Adam Yates, Karel Bezstarosti, Jeroen Demmers, Ian Chambers, and Raymond A Poot. An Oct4-centered protein interaction network in embryonic stem cells. *Cell stem cell*, 6(4):369–81, April 2010. ISSN 1875-9777. doi: 10.1016/j.stem.2010.02.014. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2860243\&tool=pmcentrez\&rendertype=abstract.

[151] Mercedes Pardo, Benjamin Lang, Lu Yu, Haydn Prosser, Allan Bradley, M Madan Babu, and Jyoti Choudhary. An expanded Oct4 interaction network: implications for stem cell

biology, development, and disease. *Cell stem cell*, 6(4):382–95, April 2010. ISSN 1875-9777. doi: 10.1016/j.stem.2010.03.004. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2860244\&tool=pmcentrez\&rendertype=abstract.

[152] Junjun Ding, Huilei Xu, Francesco Faiola, Avi Ma'ayan, and Jianlong Wang. Oct4 links multiple epigenetic pathways to the pluripotency network. *Cell research*, 22(1):155–67, January 2012. ISSN 1748-7838. doi: 10.1038/cr.2011.179. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3252465\&tool=pmcentrez\&rendertype=abstract.

[153] Zhiguang Gao, Jesse L Cox, Joshua M Gilmore, Briana D Ormsbee, Sunil K Mallanna, Michael P Washburn, and Angie Rizzino. Determination of protein interactome of transcription factor Sox2 in embryonic stem cells engineered for inducible expression of four reprogramming factors. *The Journal of biological chemistry*, 287 (14):11384–97, March 2012. ISSN 1083-351X. doi: 10.1074/jbc.M111.320143. URL http://www.ncbi.nlm.nih.gov/pubmed/22334693.

[154] Sunil K Mallanna, Briana D Ormsbee, Michelina Iacovino, Joshua M Gilmore, Jesse L Cox, Michael Kyba, Michael P Washburn, and Angie Rizzino. Proteomic analysis of Sox2-associated proteins during early stages of mouse embryonic stem cell differentiation identifies Sox21 as a novel regulator of stem cell fate. *Stem cells (Dayton, Ohio)*, 28(10):1715–27, October 2010. ISSN 1549-4918. doi: 10.1002/stem.494. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3260005\&tool=pmcentrez\&rendertype=abstract.

[155] Guang Hu, Jonghwan Kim, Qikai Xu, Yumei Leng, Stuart H Orkin, and Stephen J Elledge. A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes & development*, 23(7):837–48, April 2009. ISSN 1549-5477. doi: 10.1101/gad.1769609. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2666338\&tool=pmcentrez\&rendertype=abstract.

[156] Xi Chen, Han Xu, Ping Yuan, Fang Fang, Mikael Huss, Vinsensius B Vega, Eleanor Wong, Yuriy L Orlov, Weiwei Zhang, Jianming Jiang, Yuin-Han Loh, Hock Chuan Yeo, Zhen Xuan Yeo, Vipin Narang, Kunde Ramamoorthy Govindarajan, Bernard Leong, Atif Shahab, Yijun Ruan, Guillaume Bourque, Wing-Kin Sung, Neil D Clarke, Chia-Lin Wei, and Huck-Hui Ng. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–17, June 2008. ISSN 1097-

4172. doi: 10.1016/j.cell.2008.04.043. URL http://www.ncbi.nlm.nih.gov/pubmed/18555785.

[157] Jonghwan Kim, Jianlin Chu, Xiaohua Shen, Jianlong Wang, and Stuart H Orkin. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, 132(6): 1049–61, March 2008. ISSN 1097-4172. doi: 10.1016/j.cell.2008.02.039. URL http://www.ncbi.nlm.nih.gov/pubmed/18358816.

[158] Benjamin L Kidder, Jim Yang, and Stephen Palmer. Stat3 and c-Myc genome-wide promoter occupancy in embryonic stem cells. *PloS one*, 3 (12):e3932, January 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0003932. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2592696\&tool=pmcentrez\&rendertype=abstract.

[159] Peter B Rahl, Charles Y Lin, Amy C Seila, Ryan A Flynn, Scott McCuine, Christopher B Burge, Phillip A Sharp, and Richard A Young. c-Myc regulates transcriptional pause release. *Cell*, 141(3):432–45, April 2010. ISSN 1097-4172. doi: 10.1016/j.cell.2010.03.030. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2864022\&tool=pmcentrez\&rendertype=abstract.

[160] Charles Y Lin, Jakob Lovén, Peter B Rahl, Ronald M Paranal, Christopher B Burge, James E Bradner, Tong Ihn Lee, and Richard A Young. Transcriptional Amplification in Tumor Cells with Elevated c-Myc. *Cell*, 151(1):56–67, September 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.08.026. URL http://www.ncbi.nlm.nih.gov/pubmed/23021215.

[161] Zuqin Nie, Gangqing Hu, Gang Wei, Kairong Cui, Arito Yamane, Wolfgang Resch, Ruoning Wang, Douglas R Green, Lino Tessarollo, Rafael Casellas, Keji Zhao, and David Levens. c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells. *Cell*, 151(1):68–79, September 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.08.033. URL http://www.ncbi.nlm.nih.gov/pubmed/23021216.

[162] Eran Meshorer, Dhananjay Yellajoshula, Eric George, Peter J Scambler, David T Brown, and Tom Misteli. Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Developmental cell*, 10(1):105–16, January 2006. ISSN 1534-5807. doi: 10.1016/j.devcel.2005.10.017. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1868458\&tool=pmcentrez\&rendertype=abstract.

[163] Matthew G Guenther, Stuart S Levine, Laurie a Boyer, Rudolf Jaenisch, and Richard a Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88, July 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.05.042. URL http://www.ncbi.nlm.nih.gov/pubmed/17632057.

[164] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith A Ching, Wei Wang, Zhiping Weng, Roland D Green, Gregory E Crawford, and Bing Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–8, March 2007. ISSN 1061-4036. doi: 10.1038/ng1966. URL http://dx.doi.org/10.1038/ng1966.

[165] Tarjei S Mikkelsen, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P Koche, William Lee, Eric Mendenhall, Aisling O'Donovan, Aviva Presser, Carsten Russ, Xiaohui Xie, Alexander Meissner, Marius Wernig, Rudolf Jaenisch, Chad Nusbaum, Eric S Lander, and Bradley E Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–60, August 2007. ISSN 1476-4687. doi: 10.1038/nature06008. URL http://dx.doi.org/10.1038/nature06008.

[166] Guangjin Pan, Shulan Tian, Jeff Nie, Chuhu Yang, Victor Ruotti, Hairong Wei, Gudrun A Jonsdottir, Ron Stewart, and James A Thomson. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell stem cell*, 1(3): 299–312, September 2007. ISSN 1875-9777. doi: 10.1016/j.stem.2007.08.003. URL http://www.ncbi.nlm.nih.gov/pubmed/18371364.

[167] Xiao Dong Zhao, Xu Han, Joon Lin Chew, Jun Liu, Kuo Ping Chiu, Andre Choo, Yuriy L Orlov, Wing-Kin Sung, Atif Shahab, Vladimir a Kuznetsov, Guillaume Bourque, Steve Oh, Yijun Ruan, Huck-Hui Ng, and Chia-Lin Wei. Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell stem cell*, 1(3):286–98, September 2007. ISSN 1875-9777. doi: 10.1016/j.stem.2007.08.004. URL http://www.ncbi.nlm.nih.gov/pubmed/18371363.

[168] Martin Leeb, Diego Pasini, Maria Novatchkova, Markus Jaritz, Kristian Helin, and Anton Wutz. Polycomb complexes act redundantly to repress genomic repeats and genes. *Genes & development*, 24(3):265–76, February 2010. ISSN 1549-5477. doi: 10.1101/

gad.544410. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2811828\&tool=pmcentrez\&rendertype=abstract.

[169] Abed AlFatah Mansour, Ohad Gafni, Leehee Weinberger, Asaf Zviran, Muneef Ayyash, Yoach Rais, Vladislav Krupalnik, Mirie Zerbib, Daniela Amann-Zalcenstein, Itay Maza, Shay Geula, Sergey Viukov, Liad Holtzman, Ariel Pribluda, Eli Canaani, Shirley Horn-Saban, Ido Amit, Noa Novershtern, and Jacob H Hanna. The H3K27 demethylase Utx regulates somatic and germ cell epigenetic reprogramming. *Nature*, 488(7411):409–13, August 2012. ISSN 1476-4687. doi: 10.1038/nature11272. URL http://www.ncbi.nlm.nih.gov/pubmed/22801502.

[170] Antonio Adamo, Borja Sesé, Stephanie Boue, Julio Castaño, Ida Paramonov, Maria J Barrero, and Juan Carlos Izpisua Belmonte. LSD1 regulates the balance between self-renewal and differentiation in human embryonic stem cells. *Nature cell biology*, 13(6): 652–9, June 2011. ISSN 1476-4679. doi: 10.1038/ncb2246. URL http://www.ncbi.nlm.nih.gov/pubmed/21602794.

[171] Liangqi Xie, Carl Pelz, Wensi Wang, Amir Bashar, Olga Varlamova, Sean Shadle, and Soren Impey. KDM5B regulates embryonic stem cell self-renewal and represses cryptic intragenic transcription. *The EMBO journal*, 30(8):1473–84, April 2011. ISSN 1460-2075. doi: 10.1038/emboj.2011.91. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3102288\&tool=pmcentrez\&rendertype=abstract.

[172] Bo Wen, Hao Wu, Yoichi Shinkai, Rafael A Irizarry, and Andrew P Feinberg. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nature genetics*, 41(2):246–50, February 2009. ISSN 1546-1718. doi: 10.1038/ng.297. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2632725\&tool=pmcentrez\&rendertype=abstract.

[173] R David Hawkins, Gary C Hon, Leonard K Lee, Queminh Ngo, Ryan Lister, Mattia Pelizzola, Lee E Edsall, Samantha Kuan, Ying Luu, Sarit Klugman, Jessica Antosiewicz-Bourget, Zhen Ye, Celso Espinoza, Saurabh Agarwahl, Li Shen, Victor Ruotti, Wei Wang, Ron Stewart, James A Thomson, Joseph R Ecker, and Bing Ren. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell stem cell*, 6(5):479–91, May 2010. ISSN 1875-9777. doi: 10.1016/j.stem.2010.03.018. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2867844\&tool=pmcentrez\&rendertype=abstract.

[174] Silvina Epsztejn-Litman, Nirit Feldman, Monther Abu-Remaileh, Yoel Shufaro, Ariela Gerson, Jun Ueda, Rachel Deplus, François Fuks, Yoichi Shinkai, Howard Cedar, and Yehudit Bergman. De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. *Nature structural & molecular biology*, 15(11):1176–83, November 2008. ISSN 1545-9985. doi: 10.1038/nsmb.1476. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2581722\&tool=pmcentrez\&rendertype=abstract.

[175] Yuin-Han Loh, Weiwei Zhang, Xi Chen, Joshy George, and Huck-Hui Ng. Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells. *Genes & development*, 21(20):2545–57, October 2007. ISSN 0890-9369. doi: 10.1101/gad.1588207. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2000320\&tool=pmcentrez\&rendertype=abstract.

[176] Anh Tram Nguyen and Yi Zhang. The diverse functions of Dot1 and H3K79 methylation. *Genes & development*, 25(13):1345–58, July 2011. ISSN 1549-5477. doi: 10.1101/gad.2057811. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3134078\&tool=pmcentrez\&rendertype=abstract.

[177] Brendan Jones, Hui Su, Audesh Bhat, Hong Lei, Jeffrey Bajko, Sarah Hevi, Gretchen A Baltus, Shilpa Kadam, Huili Zhai, Reginald Valdez, Susana Gonzalo, Yi Zhang, En Li, and Taiping Chen. The histone H3K79 methyltransferase Dot1L is essential for mammalian development and heterochromatin structure. *PLoS genetics*, 4(9):e1000190, January 2008. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000190. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2527135\&tool=pmcentrez\&rendertype=abstract.

[178] Tamer T Onder, Nergis Kara, Anne Cherry, Amit U Sinha, Nan Zhu, Kathrin M Bernt, Patrick Cahan, B Ogan Marcarci, Juli Unternaehrer, Piyush B Gupta, Eric S Lander, Scott A Armstrong, and George Q Daley. Chromatin-modifying enzymes as modulators of reprogramming. *Nature*, 483(7391):598–602, March 2012. ISSN 1476-4687. doi: 10.1038/nature10953. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3501145\&tool=pmcentrez\&rendertype=abstract.

[179] Sol Efroni, Radharani Duttagupta, Jill Cheng, Hesam Dehghani, Daniel J Hoeppner, Chandravanu Dash, David P Bazett-Jones, Stuart Le Grice, Ronald D G McKay, Kenneth H Buetow, Thomas R Gingeras, Tom Misteli, and Eran

Meshorer. Global transcription in pluripotent embryonic stem cells. *Cell stem cell*, 2(5):437–47, May 2008. ISSN 1875-9777. doi: 10.1016/j.stem.2008.03.021. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2435228\&tool=pmcentrez\&rendertype=abstract.

[180] Xiangzhi Li, Li Li, Ruchi Pandey, Jung S Byun, Kevin Gardner, Zhaohui Qin, and Yali Dou. The histone acetyltransferase MOF is a key regulator of the embryonic stem cell core transcriptional network. *Cell stem cell*, 11(2):163–78, August 2012. ISSN 1875-9777. doi: 10.1016/j.stem.2012.04.023. URL http://www.ncbi.nlm.nih.gov/pubmed/22862943.

[181] Jana Krejcí, Radka Uhlírová, Gabriela Galiová, Stanislav Kozubek, Jana Smigová, and Eva Bártová. Genome-wide reduction in H3K9 acetylation during human embryonic stem cell differentiation. *Journal of cellular physiology*, 219(3):677–87, June 2009. ISSN 1097-4652. doi: 10.1002/jcp.21714. URL http://www.ncbi.nlm.nih.gov/pubmed/19202556.

[182] Shai Melcer, Hadas Hezroni, Eyal Rand, Malka Nissim-Rafinia, Arthur Skoultchi, Colin L Stewart, Michael Bustin, and Eran Meshorer. Histone modifications and lamin A regulate chromatin protein dynamics in early embryonic stem cell differentiation. *Nature communications*, 3:910, January 2012. ISSN 2041-1723. doi: 10.1038/ncomms1915. URL http://www.ncbi.nlm.nih.gov/pubmed/22713752.

[183] Carol B Ware, Linlin Wang, Brigham H Mecham, Lanlan Shen, Angelique M Nelson, Merav Bar, Deepak A Lamba, Derek S Dauphin, Brian Buckingham, Bardia Askari, Raymond Lim, Muneesh Tewari, Stanley M Gartler, Jean-Pierre Issa, Paul Pavlidis, Zhijun Duan, and C Anthony Blau. Histone deacetylase inhibition elicits an evolutionarily conserved self-renewal program in embryonic stem cells. *Cell stem cell*, 4(4):359–69, April 2009. ISSN 1875-9777. doi: 10.1016/j.stem.2009.03.001. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2719860\&tool=pmcentrez\&rendertype=abstract.

[184] Brian J Raney, Melissa S Cline, Kate R Rosenbloom, Timothy R Dreszer, Katrina Learned, Galt P Barber, Laurence R Meyer, Cricket A Sloan, Venkat S Malladi, Krishna M Roskin, Bernard B Suh, Angie S Hinrichs, Hiram Clawson, Ann S Zweig, Vanessa Kirkup, Pauline A Fujita, Brooke Rhead, Kayla E Smith, Andy Pohl, Robert M Kuhn, Donna Karolchik, David Haussler, and W James Kent. ENCODE

whole-genome data in the UCSC genome browser (2011 update). *Nucleic acids research*, 39(Database issue):D871–5, January 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1017. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013645\&tool=pmcentrez\&rendertype=abstract.

[185] Bradley E Bernstein, Ewan Birney, Ian Dunham, Eric D Green, Chris Gunter, and Michael Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. ISSN 1476-4687. doi: 10.1038/nature11247. URL http://www.ncbi.nlm.nih.gov/pubmed/22955616.

[186] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996. URL http://www.jstor.org/stable/2346178.

[187] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, January 2004. ISSN 2168-8966. URL http://projecteuclid.org/euclid.aos/1083178935.

[188] P Argos. An investigation of protein subunit and domain interfaces. *Protein engineering*, 2(2):101–13, July 1988. ISSN 0269-2139. URL http://www.ncbi.nlm.nih.gov/pubmed/3244692.

[189] S Jones and J M Thornton. Protein-protein interactions: a review of protein dimer structures. *Progress in biophysics and molecular biology*, 63(1):31–65, January 1995. ISSN 0079-6107. URL http://www.ncbi.nlm.nih.gov/pubmed/7746868.

[190] J. Janin, C. Chothia, JB Shabb, L. Ng, JD Corbin, P. B\\"utikofer, ZW Lin, DT Chiu, B. Lubin, FA Kuypers, and Others. The structure of protein-protein recognition sites. *The Journal of biological chemistry*, 265(27):16027–30, September 1990. ISSN 0021-9258. URL http://www.jbc.org/content/265/27http://www.ncbi.nlm.nih.gov/pubmed/2204619.

[191] J Janin, S Miller, and C Chothia. Surface, subunit interfaces and interior of oligomeric proteins. *Journal of molecular biology*, 204(1):155–64, November 1988. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/3216390.

[192] R A Laskowski. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of molecular graphics*, 13(5):323–30, 307–8, October 1995. ISSN 0263-7855. URL http://www.ncbi.nlm.nih.gov/pubmed/8603061.

[193] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–12, October 2004. ISSN 0192-8651. doi: 10.1002/jcc.20084. URL http://www.ncbi.nlm.nih.gov/pubmed/15264254.

[194] Andrew R Leach. *Molecular Modelling: Principles and Applications*. Pearson Education Limited, Harlow, 2 edition, 2001. ISBN 978-0-582-38210-7.

[195] C Chothia and J Janin. Principles of protein-protein recognition. *Nature*, 256(5520): 705–8, August 1975. ISSN 0028-0836. URL http://www.ncbi.nlm.nih.gov/pubmed/1153006.

[196] Irene M A Nooren and Janet M Thornton. Diversity of protein-protein interactions. *EMBO Journal*, 22(14), 2003.

[197] L Lo Conte, C Chothia, and J Janin. The atomic structure of protein-protein recognition sites. *Journal of molecular biology*, 285(5):2177–98, February 1999. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/9925793.

[198] S Miller, J Janin, A M Lesk, and C Chothia. Interior and surface of monomeric proteins. *Journal of molecular biology*, 196(3):641–56, August 1987. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/3681970.

[199] T Clackson and J A Wells. A hot spot of binding energy in a hormone-receptor interface. *Science (New York, N.Y.)*, 267(5196):383–6, January 1995. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/7529940.

[200] A A Bogan and K S Thorn. Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1):1–9, July 1998. ISSN 0022-2836. doi: 10.1006/jmbi.1998.1843. URL http://www.ncbi.nlm.nih.gov/pubmed/9653027.

[201] L Li, B Zhao, Z Cui, J Gan, M K Sakharkar, and P Kangueane. Identification of hot spot residues at protein-protein interface. *Bioinformation*, 1:121–126, 2006. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=PubMed\&dopt=Citation\&list_uids=17597870.

[202] Z Hu, B Ma, H Wolfson, and R Nussinov. Conservation of polar residues as hot spots at protein interfaces. *Proteins*, 39(4):331–42, June 2000. ISSN 0887-3585. URL http://www.ncbi.nlm.nih.gov/pubmed/10813815.

[203] Ozlem Keskin, Buyong Ma, and Ruth Nussinov. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol*, 345(5):1281–94, 2005. ISSN 0022-2836. doi: 10.1016/j.jmb.2004.10.077. URL http://www.ncbi.nlm.nih.gov/pubmed/15644221.

[204] Buyong Ma, Tal Elkayam, Haim Wolfson, and Ruth Nussinov. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10):5772–7, May 2003. ISSN 0027-8424. doi: 10.1073/pnas.1030237100. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=156276\&tool=pmcentrez\&rendertype=abstract.

[205] Inbal Halperin, Haim Wolfson, and Ruth Nussinov. Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure (London, England : 1993)*, 12(6):1027–38, June 2004. ISSN 0969-2126. doi: 10.1016/j.str.2004.04.009. URL http://www.ncbi.nlm.nih.gov/pubmed/15274922.

[206] K L Morrison and G A Weiss. Combinatorial alanine-scanning. *Current opinion in chemical biology*, 5(3):302–7, June 2001. ISSN 1367-5931. URL http://www.ncbi.nlm.nih.gov/pubmed/11479122.

[207] Garrett M. Morris, David S. Goodsell, Robert S. Halliday, Ruth Huey, William E. Hart, Richard K. Belew, and Arthur J. Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662, November 1998. ISSN 0192-8651. doi: 10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B. URL http://doi.wiley.com/10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.

[208] W D Cornell, P Cieplak, C I Bayly, I R Gould, K M Merz, D M Ferguson, D C Spellmeyer, T Fox, J W Caldwell, and P A Kollman. A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995. ISSN 00027863.

[209] Alasdair T R Laurie, Richard M Jackson, and Q-sitefinder. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics (Oxford, England)*, 21(9):1908–16, 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti315. URL http://www.ncbi.nlm.nih.gov/pubmed/15701681.

[210] D G Levitt and L J Banaszak. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of molecular graphics*, 10(4):229–34, December 1992. ISSN 0263-7855. URL http://www.ncbi.nlm.nih.gov/pubmed/1476996.

[211] M Hendlich, F Rippmann, and G Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol*, 15(6):359–63, 389, December 1997. ISSN 1093-3263. URL http://www.ncbi.nlm.nih.gov/pubmed/9704298.

[212] P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry*, 28(7):849–857, July 1985. ISSN 0022-2623. doi: 10.1021/jm00145a002. URL http://dx.doi.org/10.1021/jm00145a002.

[213] A Armon, D Graur, and N Ben-Tal. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of molecular biology*, 307(1):447–63, March 2001. ISSN 0022-2836. doi: 10.1006/jmbi.2000.4474. URL http://www.ncbi.nlm.nih.gov/pubmed/11243830.

[214] Nicholas J Burgoyne and Richard M Jackson. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics (Oxford, England)*, 22(11):1335–42, 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl079. URL http://www.ncbi.nlm.nih.gov/pubmed/16522669.

[215] T Kortemme and D Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl. Acad. Sci. USA*, 99:14116–14121, 2002.

[216] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *Journal of Molecular Biology*, 320(2):369–387, July 2002. ISSN 00222836. doi: 10.1016/S0022-2836(02)00442-4. URL http://linkinghub.elsevier.com/retrieve/pii/S0022283602004424.

[217] Y Gao, D Douguet, A Tovchigrechko, and I A Vakser. DOCKGROUND system of databases for protein recognition studies: unbound structures for docking. *Proteins*, 69:845–851, 2007. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=PubMed\&dopt=Citation\&list_uids=17803215.

[218] Solène Grosdidier, J Fernandez-recio, and J. Fernández-Recio. Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinformatics*, 9(1):447, 2008. doi: 10.1186/1471-2105-9-447. URL http://www.biomedcentral.com/1471-2105/9/447http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=PubMed\&dopt=Citation\&list_uids=18939967.

[219] Juan Fernández-Recio, Maxim Totrov, and Ruben Abagyan. Identification of protein-protein interaction sites from docking energy landscapes. *Journal of molecular biology*, 335(3):843–65, January 2004. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/14687579.

[220] K T Simons, C Kooperberg, E Huang, and D Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of molecular biology*, 268(1):209–25, April 1997. ISSN 0022-2836. doi: 10.1006/jmbi.1997.0959. URL http://dx.doi.org/10.1006/jmbi.1997.0959.

[221] Y Ofran and B Rost. ISIS: interaction sites identified from sequence. *Bioinformatics*, 23: 0–6, 2007. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=PubMed\&dopt=Citation\&list_uids=17237081.

[222] Helen M Berman, Tammy Battistuz, T N Bhat, Wolfgang F Bluhm, Philip E Bourne, Kyle Burkhardt, Zukang Feng, Gary L Gilliland, Lisa Iype, Shri Jain, Phoebe Fagan, Jessica Marvin, David Padilla, Veerasamy Ravichandran, Bohdan Schneider, Narmada Thanki, Helge Weissig, John D Westbrook, and Christine Zardecki. The Protein Data Bank. *Acta crystallographica. Section D, Biological crystallography*, 58(Pt 6 No 1):899–907, June 2002. ISSN 0907-4449. URL http://www.ncbi.nlm.nih.gov/pubmed/12037327.

[223] R M Jackson, H A Gabb, and M J Sternberg. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *Journal of molecular biology*, 276(1):265–85, February 1998. ISSN 0022-2836. doi: 10.1006/jmbi.1997.1519. URL http://www.ncbi.nlm.nih.gov/pubmed/9514726.

[224] S Campbell. Ligand binding: functional site location, similarity and docking. *Current Opinion in Structural Biology*, 13(3):389–395, June 2003. ISSN 0959440X. doi: 10.1016/S0959-440X(03)00075-7. URL http://linkinghub.elsevier.com/retrieve/pii/S0959440X03000757.

[225] P Tuffery, C Etchebest, S Hazout, and R Lavery. A new approach to the rapid determination of protein side chain conformations. *Journal of biomolecular structure & dynamics*, 8(6):1267–89, June 1991. ISSN 0739-1102. URL http://www.ncbi.nlm.nih.gov/pubmed/1892586.

[226] Marc F Lensink and Shoshana J Wodak. Blind predictions of protein interfaces by docking calculations in CAPRI. *Proteins*, 78(15):3085–95, November 2010. ISSN 1097-0134. doi: 10.1002/prot.22850. URL http://www.ncbi.nlm.nih.gov/pubmed/20839234.

[227] Marc F Lensink and Shoshana J Wodak. Docking and scoring protein interactions: CAPRI 2009. *Proteins*, 78(15):3073–84, November 2010. ISSN 1097-0134. doi: 10.1002/prot.22818. URL http://www.ncbi.nlm.nih.gov/pubmed/20806235.

[228] Joël Janin. The targets of CAPRI Rounds 13-19. *Proteins*, 78(15):3067–72, November 2010. ISSN 1097-0134. doi: 10.1002/prot.22774. URL http://www.ncbi.nlm.nih.gov/pubmed/20589643.

[229] Tanja Kortemme, David E Kim, and David Baker. Computational alanine scanning of protein-protein interfaces. *Science's STKE : signal transduction knowledge environment*, 2004(219):pl2, February 2004. ISSN 1525-8882. doi: 10.1126/stke.2192004pl2. URL http://stke.sciencemag.org/cgi/content/abstract/sigtrans;2004/219/pl2.

[230] Nurcan Tuncbag, Ozlem Keskin, and Attila Gursoy. HotPoint: hot spot prediction server for protein interfaces. *Nucleic acids research*, May 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq323. URL http://nar.oxfordjournals.org/cgi/content/abstract/gkq323v1.

[231] Nurcan Tuncbag, Attila Gursoy, and Ozlem Keskin. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics (Oxford, England)*, 25(12):1513–20, June 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp240. URL http://www.ncbi.nlm.nih.gov/pubmed/19357097.

[232] Kyu-il Cho, Dongsup Kim, and Doheon Lee. A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic acids research*, 37(8):2672–87, May 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp132. URL http://nar.oxfordjournals.org/content/37/8/2672.abstract?ijkey=f88a39d5a82d6a54d00a9cac6e83ba7c93867f99\&keytype2=tf_ipsecsha.

[233] Bin Xu, Xiaoming Wei, Lei Deng, Jihong Guan, and Shuigeng Zhou. A semi-supervised boosting SVM for predicting hot spots at protein-protein interfaces. *BMC systems biology*, 6 Suppl 2:S6, January 2012. ISSN 1752-0509. doi: 10.1186/1752-0509-6-S2-S6. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3521187\&tool=pmcentrez\&rendertype=abstract.

[234] S Velankar, C Best, B Beuth, C H Boutselakis, N Cobley, a W Sousa Da Silva, D Dimitropoulos, a Golovin, M Hirshberg, M John, E B Krissinel, R Newman, T Oldfield, a Pajon, C J Penkett, J Pineda-Castillo, G Sahni, S Sen, R Slowley, a Suarez-Uruena, J Swaminathan, G van Ginkel, W F Vranken, K Henrick, and G J Kleywegt. PDBe: Protein Data Bank in Europe. *Nucleic acids research*, pages 1–10, October 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp916. URL http://www.ncbi.nlm.nih.gov/pubmed/19858099.

[235] Richard M Jackson. Q-fit: a probabilistic method for docking molecular fragments by sampling low energy conformational space. *Journal of computer-aided molecular design*, 16(1):43–57, January 2002. ISSN 0920-654X. URL http://www.ncbi.nlm.nih.gov/pubmed/12197665.

[236] Anna Radzicka and Richard Wolfenden. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, 27(5):1664–1670, March 1988. ISSN 0006-2960. doi: 10.1021/bi00405a042. URL http://pubs.acs.org/doi/abs/10.1021/bi00405a042.

[237] K S Thorn and A A Bogan. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17:284–285, 2001.

[238] T B Fischer, K V Arunachalam, D Bailey, V Mangual, S Bakhru, R Russo, D Huang, M Paczkowski, V Lalchandani, C Ramachandra, B Ellison, S Galer, J Shapley, E Fuentes, and J Tsai. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics (Oxford, England)*, 19(11):1453–4, July 2003. ISSN 1367-4803. URL http://www.ncbi.nlm.nih.gov/pubmed/12874065.

[239] G Schreiber and A R Fersht. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *Journal of molecular biology*, 248(2):478–86, April 1995. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/7739054.

[240] G Schreiber and A R Fersht. Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. *Biochemistry*, 32(19):5145–50, May 1993. ISSN 0006-2960. URL http://www.ncbi.nlm.nih.gov/pubmed/8494892.

[241] R Wallis, K Y Leung, M J Osborne, R James, G R Moore, and C Kleanthous. Specificity in protein-protein recognition: conserved Im9 residues are the major determinants of stability in the colicin E9 DNase-Im9 complex. *Biochemistry*, 37(2):476–85, January 1998. ISSN 0006-2960. doi: 10.1021/bi971884a. URL http://www.ncbi.nlm.nih.gov/pubmed/9425068.

[242] A Ashkenazi, L G Presta, S A Marsters, T R Camerato, K A Rosenthal, B M Fendly, and D J Capon. Mapping the CD4 binding site for human immunodeficiency virus by alanine-scanning mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 87(18):7150–4, September 1990. ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=54701\&tool=pmcentrez\&rendertype=abstract.

[243] L Leder, A Llera, P M Lavoie, M I Lebedeva, H Li, R P Sékaly, G A Bohach, P J Gahr, P M Schlievert, K Karjalainen, and R A Mariuzza. A mutational analysis of the binding of staphylococcal enterotoxins B and C3 to the T cell receptor beta chain and major histocompatibility complex class II. *The Journal of experimental medicine*, 187(6):823–33, March 1998. ISSN 0022-1007. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2212189\&tool=pmcentrez\&rendertype=abstract.

[244] A Rajpal, M G Taylor, and J F Kirsch. Quantitative evaluation of the chicken lysozyme epitope in the HyHEL-10 Fab complex: free energies and kinetics. *Protein science : a publication of the Protein Society*, 7(9):1868–74, September 1998. ISSN 0961-8368. doi: 10.1002/pro.5560070903. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2144172\&tool=pmcentrez\&rendertype=abstract.

[245] H Hwang, B Pierce, J Mintseris, J Janin, and Z Weng. Protein-protein docking benchmark version 3.0. *Proteins*, 73:705–709, 2008. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=PubMed\&dopt=Citation\&list_uids=18491384.

[246] HR Guy. Amino acid side-chain partition energies and distribution of residues in soluble proteins. *Biophysical journal*, 47(1):61–70, 1985. URL http://linkinghub.elsevier.com/retrieve/pii/S0006349585838777.

[247] R Wolfenden, L Andersson, P M Cullis, and C C B Southgate. Affinities of amino acid side-chains for solvent water. *Biochemistry*, 20:849–855, 1981.

[248] L Wesson and D Eisenberg. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein science : a publication of the Protein Society*, 1(2): 227–35, February 1992. ISSN 0961-8368. doi: $10.1002/\mathrm{pro}.5560010204$. URL `http://www.ncbi.nlm.nih.gov/pubmed/1304905`.