

Analysis of Transcription Factor Binding at *cis*-Regulatory Elements during Blood Development and Differentiation

Maria Cristina Suci

St Cross College

**MRC Molecular Haematology Unit
Weatherall Institute of Molecular Medicine**



*Thesis submitted for the degree of Doctor of Philosophy
University of Oxford*

Trinity 2015

*To my parents,
For their unconditioned support.*

*Parintilor mei,
Pentru sprijinul neconditionat.*

I. Abstract

The overall aim of this work is to understand how regulatory sequences (enhancers, promoters) are bound and interpreted by TFs during erythroid development and differentiation. This is becoming increasingly important as DNA sequence variants in enhancers may alter protein binding and consequently may contribute to changes in gene expression that affect susceptibility to common diseases.

We hypothesised that by using high-resolution analysis of chromatin accessibility data, we would identify relevant differences in sequence usage at regulatory regions in erythroid cells at various stages of development and differentiation.

To test this, we developed a robust and efficient method to isolate HS regions in erythroid cells by performing DNase-Seq and ATAC-Seq. When analysed at high resolution, open-chromatin assays, namely DNase-Seq, reveal a genome-wide profile of “footprints” - transcription factor bound sites at regulatory elements in different erythroid cell types.

The two dynamic contexts, development and differentiation, in which we analysed the changes in HS regions share a common pattern: we identified approximately 10% cell specific HS regions. The differential open-chromatin regions correspond mainly to enhancers, but also promoters. Majority of these elements (~80%) occur in different genomic regions. These observations reinforce the fact that differentiation and development are driven by TF binding at distal elements.

At the α -globin locus, the quantitative changes of HS regions are correlated with the activity of gene promoters. The level of HS enrichment refines the chromatin signature of functional enhancers.

Novel analysis of HS genome-wide data, “average footprinting”, provides a constrained set of transcription factor binding sites ranked by their usage in a specific cell type. We also adapted this approach to predict the effect of a variant in a sequence on altering affinity to TF and potentially causing changes in gene expression. This predicted damage score approach can be used to prioritise variants for functional analysis.

Finally, we have identified families with unexplained anaemia, which have a SNP in a binding site of the major enhancer of the human α -globin cluster. This variant has a high damage score based on our prediction analysis. We are currently investigating if, and how, this non-coding SNP affects gene expression as an example of how regulatory SNPs in general may contribute to human genetic disease.

Acknowledgements

I would like to thank my supervisors Prof Doug Higgs and Prof Jim Hughes for the support and advice throughout my project. They have helped me maximise my potential by offering a stimulating environment in which to enrich not only my scientific skills but also my critical thinking.

To Dr Mira Kassouf, my mentor and my friend, who has offered so much encouragement and needed guidance. I am extremely grateful and inspired. This process was made not only bearable, but enjoyable because of the invigorating discussions and clarity she brought into my scientific world (and not only). An inspiration.

To Dr Bryony Graham and Dr Hsiao Voon, who have always been there to provide an answer to my unlimited (silly) questions. They taught me patience and how to be resilient during the scientific marathon. To all the other members of the Higgs, Gibbons labs and members of the MHU unit, especially Ms Helena Ayyub, who patiently introduced me to the 'forgotten' art of molecular biology classics, such as Southern blotting.

My days in the MHU would not have been as joyful without my dear friend Chrissy. She was there for the small victories, and the big ones- from PCRs to Confirmation. Together, we would shrug off any scientific defeat with a giggle. My true partner in 'crime'.

My DPhil years have been marked by meeting great people who I was lucky enough to have not only as housemates but also as friends, my 'PhD' family: Carme, Kiran, Robbie, Roos and Na. Our shared nerdiness that could turn any meal into a complex scientific nonsense, quirky dinners (of course no peanuts or tomatoes) and awesome parties, turned Oxford into 'home'. My dear friend Sonia, though a bit further away now than a college block, has been there to listen to all things good and bad about Oxford, science and life in general. Her friendship, and experience have been invaluable.

To Anthony, who has endured through these past few months of creation, and facilitated the arduous process of thesis writing by providing love, laughter, chocolate and an indispensable white board. And probably a very required dose of sanity. Thank you for always being there.

Finally, none of this would have been possible without the support of my family: my mother Corina, my father Sorin and my brother Nicolae. My gratitude cannot be put into words. My parents have nurtured my insatiable curiosity and taught me how to question everything around. They pushed me outside my comfort zone, and inspired me to follow my dreams. And when those dreams involved leaving for the big world, they supported me through hardships only they know, so I can study and have a career I truly enjoy. It is only because of their unlimited and unconditioned support that I made it all this way. All my love goes to them, my role models, my inspiration.

Aceasta realizare nu ar fi fost posibila fara familia mea: parintii mei Corina si Sorin, si fratele meu Nicolae. Cuvintele nu sunt de ajuns pentru a-mi exprima recunostiinta. Parintii mei au fost cei care mi-au cultivat curiozitatea nelimitata, si m-au invatat sa gandesc critic. M-au invatat sa vreau mai mult, sa imi depasesc limitele, si sa imi urmez visele chiar daca drumul nu a fost usor. Cand visele au insemnat sa plec departe de casa, m-au sprijinit impotriva nenumaratelor greutati pentru a invata si a face ceea ce-mi place. Parcursul meu si reusita de a-mi obtine doctoratul se datoreaza dragostei si incurajarilor lor nelimitate .

Abbreviations

3C	chromosome conformation capture
Ab	antibody
APC	allophycocyanin
APC-Cy7	allophycocyanin Cy7
APH	acetylphenylhydrazine
ATAC	Assay for transposition of native chromatin
BFU-E	burst forming unit-erythroid
bHLH	basic helix-loop-helix
cDNA	complementary DNA
°C	degrees Celsius
ChIP	chromatin immunoprecipitation
CFU-E	colony forming unit-erythroid
CRISPR	clustered regularly interspaced short palindromic repeats
DMEM	Dulbecco's modified Eagle's medium
DMSO	dimethylsulphoxide
DNA	deoxyribonucleic acid
DNaseI	Deoxyribonuclease I
dNTP	2'-deoxynucleotides 5'-triphosphate
EKLF	erythroid Krüppel-like factor
EPO	erythropoietin
eQTL	expression quantitative trait loci
EryD	definitive erythroid cells
EryP	primitive erythroid cells
ES	embryonic stem cells
FACS	fluorescence activated cell sorting
FCS	fetal calf serum
FITC	fluorescein isothiocyanate
FOG-1	friend of Gata 1
FSC	forward scatter
G1, G2	gap 1, gap 2
G1E	GATA1 erythroid

GWAS	genome-wide association study
HS	hypersensitive
HS regions	hypersensitive regions
HGB	haemoglobin
HSC	haematopoietic stem
IF	immunofluorescence
Kb	kilobase
KO	knock-out
MCH	mean corpuscular haemoglobin
MCV	mean corpuscular volume
MEL	mouse erythroleukaemia
mg	milligram
MGG	May-Grünwald Giemsa
min	minute
ml	millilitre
mM	millimolar
mRNA	messenger RNA
NGS	next generation sequencing
P/S	penicillin/streptomycin
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PI	propidium iodide
RNA	ribonucleic acid
RT	room temperature
RT-PCR	reverse transcription-PCR
SCL	stem cell leukaemia
Seq	sequencing
SSC	side scatter
U	unit
UV	ultraviolet
WT	wildtype
YS	yolk sac
µg	microgram
µl	microlitre

Table of Contents

I. ABSTRACT	I
I. INTRODUCTION.....	1
1.01 <i>Eukaryotic gene regulation</i>	3
1.02 <i>Regulatory sequences - platforms for gene regulation</i>	5
1.03 <i>Methods for assaying protein-DNA interactions at basepair resolution</i>	10
1.04 <i>Variants for common disease concentrate in regulatory regions.....</i>	17
1.05 <i>The case for model loci in an era of genome-wide analyses.....</i>	19
1.06 <i>Erythropoiesis, model system for molecular genetics.....</i>	28
1.07 <i>Differentiation of definitive red blood cells</i>	36
AIM	38
II. MATERIALS AND METHODS	40
1.01 <i>Isolation of cell populations.....</i>	41
1.02 <i>Quantitative Real-Time PCR</i>	45
1.03 <i>Cellular staining</i>	46
1.04 <i>Embryonic fetal liver progenitors assay.....</i>	47
1.05 <i>Open-chromatin assays.....</i>	48
1.06 <i>Naked DNA controls for DNase-Seq and ATAC-Seq</i>	51
1.07 <i>Sequencing.....</i>	51
1.08 <i>Southern Blot.....</i>	53
1.09 <i>Computational analysis.....</i>	53
III. RESULTS	58
CHAPTER I. OPEN-CHROMATIN ASSAYS: OPTIMISATION AND SETUP	59
1.02 <i>Assessing hypersensitivity using the DNaseI assay.....</i>	61
1.03 <i>Digital footprinting with DNase-Seq and ATAC-Seq data.....</i>	104
CHAPTER II. ANALYSIS OF CIS-REGULATORY ELEMENTS IN ERYTHROID DEVELOPMENT	114
2.01 <i>Introduction.....</i>	114
2.02 <i>Chromatin accessibility at the α-globin cluster in primitive and definitive erythroid cells</i>	118
2.03 <i>Genome-wide analysis</i>	128
2.04 <i>Conclusion</i>	137
CHAPTER III. ANALYSIS OF CIS-REGULATORY ELEMENTS IN ERYTHROID DIFFERENTIATION	139
3.02 <i>Purification of primary erythroid populations corresponding to two different stages of erythropoiesis</i>	141
3.03 <i>Chromatin features of the $\alpha\alpha$-globin cluster differ between the two stages of differentiation</i>	149
3.04 <i>Genome-wide analysis of HS of at the two stages of differentiation.....</i>	159

3.05 Conclusion	169
CHAPTER IV. A NOVEL APPROACH FOR THE ANALYSIS OF FOOTPRINTS	170
4.01 Introduction	170
4.02 Computing genome-wide footprints - ‘Average’ footprints from DNase-Seq data 171	
4.03 Computing “average” footprints from ATAC-Seq data.....	187
4.04 Conclusion	193
CHAPTER V. ANALYSIS OF VARIANTS IN-REGULATORY SEQUENCES IN HUMAN ERYTHROID CELLS 194	
5.01 Introduction	194
5.02 HS regions at the human α -globin cluster	197
5.03 Using “average” footprints to assess impact of variants in regulatory sequences 205	
5.04 Conclusion	223
IV. DISCUSSION	224
SUMMARY.....	225
Isolating regulatory regions- Open chromatin assays.....	227
α -globin cluster at different developmental stages- Primitive and Definitive cells.....	233
α -globin cluster at different stages of erythropoiesis - early progenitors and differentiated cells	236
Protein-DNA interactions- Factors involved in TF binding to regulatory regions	243
Determining the role of variation in disease- challenges and solutions	259
V. APPENDICES AND REFERENCES	268

I. Introduction

“The central problem in biology was how the one-dimensional sequence of nucleotides in DNA specified the one-dimensional sequence of amino acids in proteins.” Sydney Brenner

While this is true for the coding sequence, and made the study of a linear, one dimensional sequence of nucleotides manageable once the genetic code was revealed, coding sequence represents only 2% of the genome ¹. The rest of 98% of sequence is non-coding and does not abide by the same rules. It is probably fair to say that one of the central problems in biology is now understanding how the non-coding genome (or “junk” as it has been called) achieves tissue-, spatial- and temporal-specific gene regulation, thus driving organism complexity.

As we zoom in, through tightly packed chromatin, modified by histones and wrapped around nucleosomes, we find accessible regions where the sequence, even though non-coding, is still of interest. Within *cis*-regulatory elements such as enhancers, this linear sequence serves as a platform for highly orchestrated binding of tissue specific transcription factors (TF), process at the heart of gene regulation.

The information that determines TF binding is encoded in the sequence, but we do not yet understand the rules of the TF binding “grammar” or all the factors involved. As an effort to unravel such rules, over-represented sequences in regulatory regions have been analysed with limited success due to the inability to discriminate bound from unbound sites. One of the major limitations is the resolution of current methods to visualise and assess TF binding. In this project, we will address this problem by studying the marks (footprints) left by TFs at regulatory sequences, to gain insight into the rules driving such orchestrated binding.

1.01 Eukaryotic gene regulation

Increasing developmental complexity in multicellular organisms (for example *C.Elegans* compared to humans) is thought not to depend on larger gene number, but on highly regulated expression patterns and alternative isoforms. Transcriptional regulation is required to ensure the appropriate genes are switched on and off at the right stage in development and differentiation.

Gene regulation in eukaryotes is the result of interplay between chromatin structure, transacting factors, additional co-factors and their interactions in the three dimensional space. In the nucleus, DNA is tightly packed around histones (nucleosomes) and condensed into chromatin (Figure 1). The more compacted the chromatin is, the less accessible it is to protein binding and, thus becomes the less active. Covalent modifications influence chromatin condensation and are associated with varying degrees of DNA accessibility. When in an open-chromatin conformation, DNA can be bound by transcription factors (TFs), and co-factors. Though the causal relationship between chromatin modifications and open-chromatin has not been established, it is clear that increased accessibility of regulatory sequences leads to TFs binding and chromatin interactions leading to genes transcription²⁻⁴. These regulatory sequences are found within cis regulatory elements, or *cis*-regulatory modules (enhancers, promoters, silencers, insulators) and encode the rules governing orchestrated gene expression patterns required for correct development⁵⁻⁷.

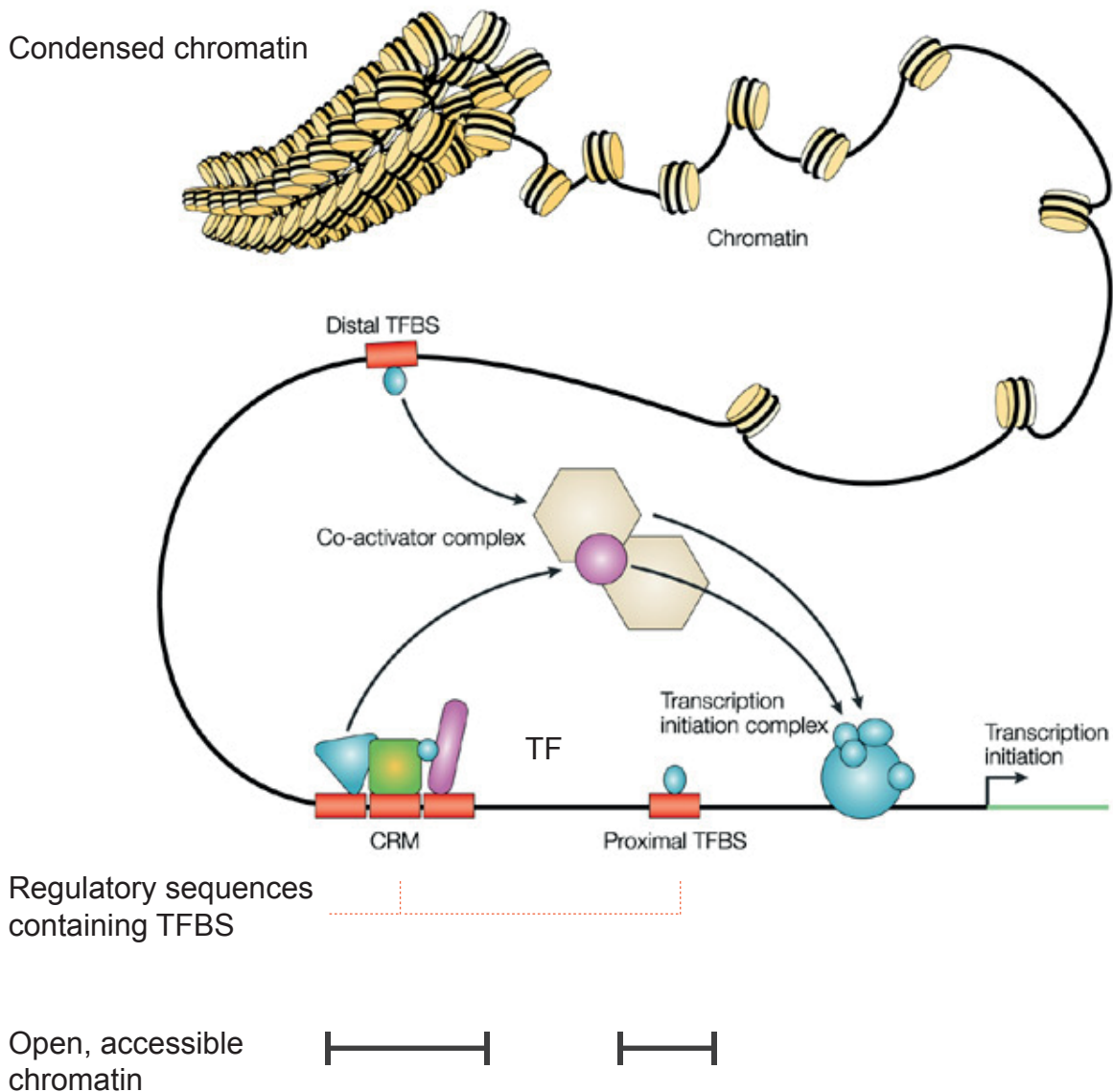


Figure 1. Factors involved in regulation of eukaryotic transcription. Within *cis*-regulatory modules (CRM), transcription factors binding sites (TFBS) are recognised by tissue specific transcription factors (TFs). These regulatory regions are found in accessible chromatin, at varying distances from a transcription start site. Bound TFs recruit co-activators and form complexes that stabilise the transcription-initiation machinery to enable gene expression. The chromatin state, influenced by histone modifications, plays an essential role in the regulation process, as TFs can only bind to open chromatin. This figure was adapted from (Williams and Sandelin, 2004)

1.02 Regulatory sequences - platforms for gene regulation

Decoding the regulatory sequences and understanding how information is converted into meaningful transcriptional signals holds the promise of not only improving our understanding of gene regulation, but also of disease. Within regulatory sequences lie the instructions for binding of regulatory proteins (transcription factors, TFs), recruitment of the transcriptional machinery, transcription initiation and robust expression in a timely and tissue specific manner. It is not surprising that alterations of the regulatory sequences have been linked to cancer^{8,9} or common disease in GWAS¹⁰⁻¹².

Despite major progress in recent years in isolating and measuring protein or TF binding to regulatory sequences in one form or another (*in vitro* protein studies or *in vivo* genomics studies), or characterising their potential to drive expression, we do not fully understand the rules governing recognition and TF binding to these regulatory sequences¹³.

The difficulties in understanding regulatory sequences 'grammar' are due to features of these sequences: (1) most often they lie in non-coding regions; (2) they are interpreted by tissue specific TFs and thus tissue specific data is necessary; (3) current methods do not have the required base-pair resolution; (4) the number of other factors (i.e. co-factors) involved make context essential.

The building blocks of regulatory sequence, TF binding sites, are subject to orientation and spacing rules to ensure the correct binding of TFs and regulate gene expression. We will focus mainly on the TF binding sites at regulatory sequences

within enhancers, as enhancers play key roles in fine-tuning gene transcription by alternating activation and silencing of specific genes in a time-dependent fashion, thus driving development from embryonic to adult life and maintenance thereafter.

(a) Different modes of TF binding have different enhancer output

Different modes of TF binding sites occupancy result in different transcriptional outputs and play an important role in the regulation of enhancer activity. In an additive binding model, enhancer activation is proportional to the concentration of individual TFs bound. Such an example is the nuclear factor κ B (NF- κ B) for which increased concentrations result in gradual increase of transcription at target genes, as a response to stimuli intensity¹⁴. As opposed to a graded output, a cooperative binding model shows no linear relationship between TF concentration and binding at enhancer elements, allowing the TFs to function at different concentrations. The first example for such a binary mode of enhancer activity was first described in phage- λ cl repressor¹⁵. This is most likely the model found in developmental contexts, where a switch mode (“on”/“off”) of enhancer action is desired. The developing embryo can be thus protected from variations in individual TFs levels, as long as the relevant TF reaches the desired threshold¹⁶.

Cooperative binding of TFs at enhancers and formation of regulatory complexes is dependent on the interplay between sequence and protein-protein interactions. This model of binding is susceptible to changes in motif ‘grammar’ and thus leading to changes in enhancer behaviour.

A well-studied example of regulatory sequence constraint in terms of motif positioning and spacing is the interferon- β enhancer, bound cooperatively by eight TFs. Any alteration of the underlying 55bp sequence affects the binding of all eight TFs^{17,18}. These observations gave rise to the model of 'enhancesome', probably an example of extreme cooperative binding due to the role of interferon- β in the rapid response to viral infections.

A more flexible model of is the 'billboard' model in which TFs bind cooperatively, but also in an additive or independent manner. The motif grammar within these type of regulatory sequences is flexible as compared to the 'enhancesome'. This is probably the likely model for the majority of developmental enhancers, as it allows for flexibility in the way the TF binding sites are occupied^{16,19}.

A third model is the 'TF collective' in which dependence on the regulatory sequence appears to be low and protein interactions seem to play the main role. This model was described in cardiac development: a group of TFs bind and co-regulate a set of enhancers without any shared motif grammar. If one of the TFs is removed, the remaining TFs fail to activate the enhancers suggesting that the overall protein-protein interactions are a stronger determinant of TF recruitment and enhancer activity than the individual TF binding sites²⁰.

Coordinate action has not only been described in relation to TF binding at individual enhancers, but also when describing classes of enhancers located in close proximity.

(b) Other factors impacting on enhancers activity

The ability of a group of core enhancers, and flanking elements, to confer high-level, tissue-specific expression on linked genes has led to defining the locus control regions (LCR)²¹ in the early days of gene regulation, and to 'super-enhancers' more recently²². The overlap in the definitions of such entities of enhancers is evident. It is not yet clear if the elements of the LCRs, or 'super-enhancers' have a synergistic function, representing more than the sum of functions of the individual enhancer components.

This topic is beyond the scope of this thesis but is worth mentioning there are assumed enhancer classes that may affect the way regulatory elements modulate gene expression. Some of the best-studied regulatory regions classified as LCRs or super-enhancers are the enhancers upstream of the α - and β - globin genes.

Another important facet of enhancer function is the physical interaction of enhancers with their linked promoter. The frequency of contacts between enhancers and promoters can be assessed by chromosome conformation capture (3C) technologies. Many variants of these assays have been developed which can assess contacts at different resolutions and scale and are applicable to different questions: (3C)- one-to-one contacts between selected genomic regions; (4C)-one-to-all; (CaptureC)- selected many-to-all; (5C)-many-to-many; (Hi-C)all-to-all²³. The field has progressed from viewing the interactions as looping between two distinct genomic regions, to a more inclusive view of a compartment of interactions, or even broad chromosomal domains, called topological domains (TAD)²⁴. The determinants of the specificity of these enhancer-promoter interactions as well as other mechanistic details remain to

be elucidated. However, it is clear from these studies that the interaction landscape is highly complex, and the assumption that an enhancer regulates the closest gene is likely to be very unreliable.

This aspect of enhancer activity is critical to have in mind as in order to study the effect of direct binding of TFs at specific enhancers on their function, we must be able to assess the activity of the promoters they regulate. Thus, we want to investigate enhancers where the link with the target promoter has been demonstrated.

1.03 Methods for assaying protein-DNA interactions at basepair resolution

Dissecting the regulatory sequence, and their component TF binding sites is complicated by the multiple factors involved in the process of gene regulations and by the limitations of the available technologies in term of resolution at which sites can be investigate.

Recent developments of high-throughput methods have allowed for improved characterisation of TF binding sites. Two main approaches, in two different fields, are used. (1) In proteomics, a more quantitative approach to measure the *in vitro* affinities of TFs to short sequences uses protein-based high-throughput assays. Though informative, discrepancies between affinities observed *in vitro* and preferred binding sites *in vivo* have been reported¹³. (2) In genomics, qualitatively measuring binding preferences along the genome by capturing *in vivo* binding events through methods such as chromatin immunoprecipitation (ChIP-chip, ChIP-Seq), or open-chromatin assays (DNase-Seq, ATAC-Seq). These methods provide us with extensive information from *in vivo* binding, but are not quantitative in terms of protein affinities.

Due to our interest to study gene regulation by regulatory regions *in vivo*, genome-wide, and the purpose of this thesis, we will focus on the second type of technologies.

(a) Chromatin immunoprecipitation assay followed by sequencing, ChIP-Seq

The chromatin immunoprecipitation (ChIP) assay is a high-throughout method for identification of genome-wide regions bound *in vivo* by a protein of interest. Proteins are crosslinked to DNA and the fragments bound by the protein of interest are enriched for by using specific antibodies. These DNA fragments are finally precipitated and can be identified by sequencing ²⁵.

The power of Chip-Seq assays lies in the fact that the protocol is applicable in various cell types, and one can identify genome-wide target sites of a TF or histone mark of interest. Limitations lie in the dependency on efficient antibodies, inability in some TF cases to distinguish direct from indirect binding or resolution. Though improved methods for ChIP resolution have been published (ChIP-exo, ChIP-nexus) ^{26,27}, the protocols are more challenging than conventional ChIP-Seq and are not yet widely used. Finally, ChIP-Seq only allows for the investigation of known factors and can reveal binding of a single protein/chromatin mark per experiment, making it quite a biased approach.

(b) Open-chromatin assays

(i) DNaseI hypersensitive sites sequencing, DNase-Seq

DNaseI hypersensitive site sequencing (DNase-Seq) is the high-throughput adaptation of the DNaseI assay used for identifying open-chromatin regions. TF binding to DNA disrupts the nucleosomal structure and results in exposure of accessible DNA, susceptible to the non-specific cutting of the DNaseI

endonuclease. Under assay conditions, DNaseI digestion takes place preferentially at accessible DNA rather than at DNA wrapped around nucleosomes. These accessible regions, found at regulatory elements in the genome, were termed hypersensitive sites (HS) and have been classically characterised using a southern blot strategy. While suitable for interrogation of single sites, such as enhancers and promoters of known loci²⁸, the DNaseI assay is not sufficient for identifying regulatory sequences genome-wide.

This has changed with the advent of genomics, when the DNaseI assay was adapted for high-throughput in various forms: (1) using biotinylated linkers and hybridisation to a tiled microarray²⁹ or sequencing^{30,31}. This method showed variable sensitivity in detecting well-characterised HSs²; (2) using size-fractionation of the digested material through sucrose-gradients^{32,33} or library size selection³⁴ and sequencing. Though slightly different, the last two methods have a similar approach which proved highly sensitive and specific and resulted in them being applied across cell types in projects such as ENCODE^{33,35}.

Unlike ChIP-Seq, DNase-Seq is TF-independent and does not require antibodies, making it possible in one experiment to identify most TF bound regions. The sequenced fragments are processed in the same manner as ChIP-Seq data, and the output has a similar pattern of peaks in areas which are enriched. Another notable difference occurs at the analysis stage as the number of peaks identified from a DNase-Seq (HS regions) can be several orders of magnitude higher than the number of ChIP-Seq peaks, as they include most, if not all, the regulatory regions bound by proteins in a cell type. Thus, depending on the peak calling stringency, the numbers can range from 20000^{34,36} to 250000 peaks^{33,37}.

Even though isolating open-chromatin is the desired method to identify regulatory elements genome-wide, DNase-Seq has limitations that can make it prohibitive to use: it requires high amounts of starting material (5×10^7 cells), it is costly because of the recommended sequencing depth to detect enrichment and the protocol needs optimisation on a cell type specific basis due to chromatin packaging characteristics. The HS peaks have similarly low resolution as the ChIP-Seq peaks, covering 250-300 bp and thus make it hard to pinpoint the TF binding sites that are bound by proteins. One major advantage of DNase-Seq is that with appropriate computational tools, the resolution issue can be solved by generation of DNaseI footprints (explained below). Finally, one DNase-Seq experiment reveals all the genomic regions that are bound by any protein.

(ii) Improving on resolution- from DNase-Seq to DNase-Seq digital footprints

DNaseI digital footprinting³⁸ is based on the principle that, within a DNaseI HS, proteins bound to DNA will protect it from DNaseI digestion. If sequenced at sufficient depth, the frequency of DNaseI cutting can be used to identify the regions ('footprints') where a TF is directly bound to DNA^{33,38}. The concept that protein binding protects the underlying DNA from endonuclease activity is not new at all. In fact DNase footprinting was described more than 30 years ago^{39,40} and since has been used to characterise specific regulatory elements through variations such *in vitro* footprinting⁴¹ or DMS mediated *in vivo* footprinting^{42,43}.

The advantage of DNase-Seq digital footprinting is the ability to identify the footprints in the same data set used to visualise DNaseI HS, meaning of course genome-wide footprints. However, the generation of such footprints is limited by

cost of sequencing for the necessary coverage, and the availability of computational tools to visualise and identify such footprints.

(iii) Assay for transposition of active chromatin followed by sequencing, ATAC-Seq

ATAC-Seq is a method for isolating open chromatin regions using modified Tn5 transposase which can insert sequencing adapters into the native chromatin. The assay is a rapid two-step protocol and can use input as low as 500-50000 cells⁴⁴. The initial part of the protocol is similar to DNase-Seq in terms of nuclear isolation. The nuclei are subsequently incubated with the Tn5 transposase, which cuts the open chromatin regions and inserts adapters, all in one preparation step. The fragments are purified, and as they already have adapters ligated, can be sequenced without further processing.

Data analysis is identical to DNase-Seq data processing and, as will be discussed in Results-Chapter 1, the enrichment of open-chromatin from ATAC-Seq is similar if not better. The authors of the ATAC-Seq study claim that this data is also suitable for generation of footprints. We discuss ATAC-Seq footprints at the end of Results-Chapter 1.

(iv) Digital Footprinting -Mechanistic insights

Footprinting analysis allows for visualisation at base-pair resolution of protein-DNA interaction. This is, without doubt, very appealing as we can progress from predicting regions that are bound by sequence analysis to identifying those regions by the footprint left on the DNA.

One of the main advantages of such analysis is the ability to identify the directly bound DNA regions. One can imagine a scenario where within a hypersensitive site that contains several motifs, we observe ChIP-Seq enrichment for multiple TFs. By adding the footprinting information, we can discriminate between directly and indirectly bound TFs and will reveal variants located within bound sites (Figure 2).

In terms of limitations, the footprints will not provide any insight into the changes caused by protein-protein interactions, and indirect TF binding. In addition there is very little known about the affinity of the proteins that form a footprint and the biochemical properties of protein binding without footprints (i.e. do highly dynamic proteins form footprints?). Thus, while the presence of a footprint is indicative of protein-DNA binding, most often supported by the underlying motif, the absence of a footprint may refer to either of absence of protein binding or the inability to detect a footprint with current methods.

Being able to identify the regions directly bound by TF is essential for deciphering the first layer of transcription regulation complexity. Moreover, the digital footprints can offer a direct assessment of sequence variation and its effect on TF binding. Analysing sequence variants located within footprints increases the chances of finding causal variants for common disease^{13,45} (discussed more in detail below).

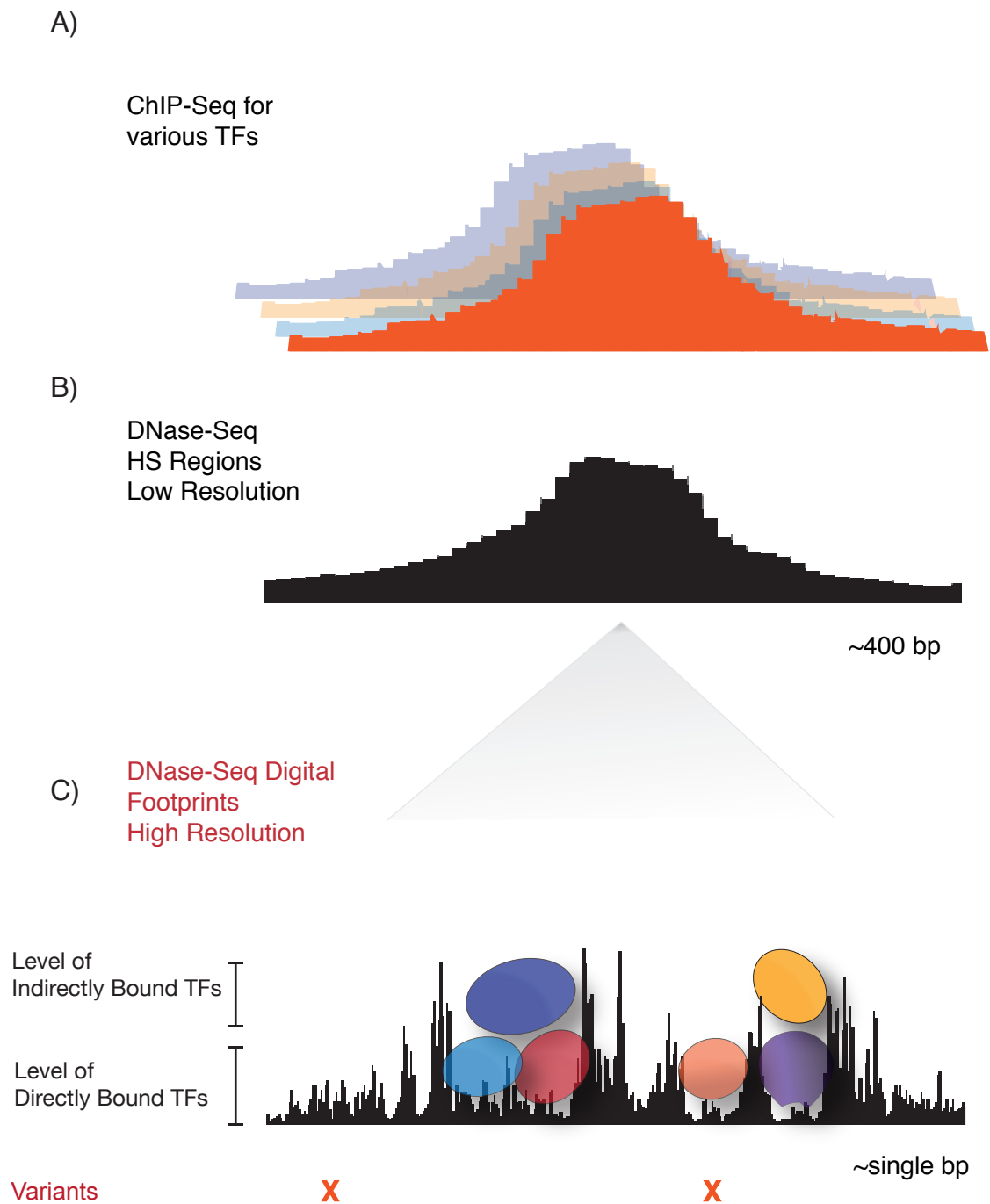


Figure 2. DNase-Seq footprints reveal binding at base-pair resolution. Diagram illustrating a regulatory sequence and the resolution of current methods. A) ChIP-Seq profiles reveal multiple factors bound at the region. B) The region is hypersensitive as shown by DNase-Seq. C) DNase-Seq footprinting reveals the sequences that are directly bound within a hypersensitive site, and potential sequence variants.

1.04 Variants for common disease concentrate in regulatory regions

Considering the surge in the last decade in available genetic information as well as phenotype data from large numbers of individuals, major efforts have been made to understand the genetic variation underlying common disease.

In genome-wide association studies (GWAS), the frequencies of different genotypes and single nucleotide polymorphisms (SNPs) in respect to the studied phenotype are analysed to obtain a set of significant associations.

The observation that the vast majority of associated SNPs (94%) are found in the non-coding regions^{11,46}, highlighted the challenges of studying variants that do not have a direct effect (by mutating the coding sequence and thus the protein sequence), but rather an indirect one by affecting gene regulation.

The bottleneck in GWAS is identifying the causal SNPs from the hundreds of non-coding associations. The desired approach would be to functionally validate the associated variants, but this is not trivial because of: (1) linkage disequilibrium (LD)-the causal variant could be a SNP in LD with the associated one; (2) the tissue/ cell type the variant affects is unknown: the associated variants are identified from genomic DNA, independent of the tissue in which the SNPs might have an effect.

The GWAS field is addressing these challenges through analyses such as expression quantitative trait loci (eQTL): the study of the variants in relation to gene expression. An eQTL analysis is based on the fact that variants can explain the levels of a transcript; thus, the transcript abundance can be considered a

quantitative trait locus which can be mapped with high statistical power and can be used to test direct association of genetic variation and gene expression levels in thousands of individuals^{46,47}.

The contribution from the gene regulation field is clear, as the field can add the functional validation through the progress made on understanding the mechanisms of distal regulation, and the defects leading to diseases^{11,12,48}. The extensive annotation and characterisation of the regulatory regions in the non-coding genome (enhancers, promoters, silencers, insulators) can only clarify the effect sequence variation might have on these regions and their activity.

The combination between the information from the GWAS studies and the tools used by the gene regulation field to annotate regulatory regions holds promise for understanding susceptibility to common disease.

It has been suggested by studying ENCODE data that mapping of TF sites within a HS using DNase-Seq footprinting can add confidence to the prediction of causal variants⁴⁹. We take this suggestion one step further and in the present thesis illustrate how analysis of open-chromatin at regulatory sequences can be used to prioritise non-coding variants for functional study.

1.05 The case for model loci in an era of genome-wide analyses

The possibility to study gene regulation processes and the factors involved, in a genome-wide fashion, has been celebrated by the community and proves revealing in many aspects such as: most of the genome is non-coding and plays a role in gene regulation, chromatin does not serve only packaging purposes but is involved in transcription through histone modifications, majority of common disease variants are in non-coding regions etc. In addition, the genomics era resulted in the generation of a wealth of data, facilitated by technological advancements especially in the field of sequencing.

However, in order to make sense of these large amounts of information, and be able to go beyond the surface of associations to functional significance, there is a need for mechanistic insights. We would argue, that there is no better way to explore genetic mechanisms than at model loci, where the variables involved are well-characterised and thus the phenotypes can be explained to a larger extent than elsewhere in the genome.

In the current thesis, we want to gain insight into the high-resolution profiles of HS regions so we can analyse sites directly bound by TFs. We plan to address these issues both at a model locus (α -globin) and at a genome-wide level.

Model loci, such as the globins, are paradigms of gene regulations and offer the advantages of being extensively studied. Challenges like annotating the globins regulatory regions, and linking to the promoters (very current topics at a genome-wide level) have long been overcome through the study of natural mutations and

knockout models. This allows for testing of more detailed aspects of distal regulation such as how are the regulatory sequences bound and interpreted by TFs, how is the hypersensitivity associated with activity or what are the TF dynamics in differentiation and development. Of course, these observations can finally be explored at a genome-wide level to derive general principles, or understand locus-specific features.

(a) The α -globin cluster in the mouse

The α -globin locus is highly conserved and contains a 135-155 kb region of synteny maintained in 22 divergent species of vertebrates studied (including mammals, birds, fish), spanning 500 million years of evolution^{50,51}. This syntenic region consists of the α -globin cluster, its regulatory elements (four in the human, conserved, and five in the mouse) and a number of genes lying upstream.

The murine α -globin cluster (Figure 3) is located on chromosome 11 and includes an embryonic α -like gene (ζ -globin), duplicated α and θ adult genes, and a pseudo-gene $\psi\alpha$ ⁵⁰. Transcription of the α -globin gene is regulated by a group of erythroid specific enhancers: four intragenic, located in the Nprl3 gene, and one intergenic located upstream of the Nprl3 promoter^{52,53}. Four of the regulatory elements (R1 to R4) are conserved across the 22 species, while one is mouse specific (Rm) . R2, corresponding to HS-40 in human, has the highest degree of conservation⁵¹ (Figure 3).

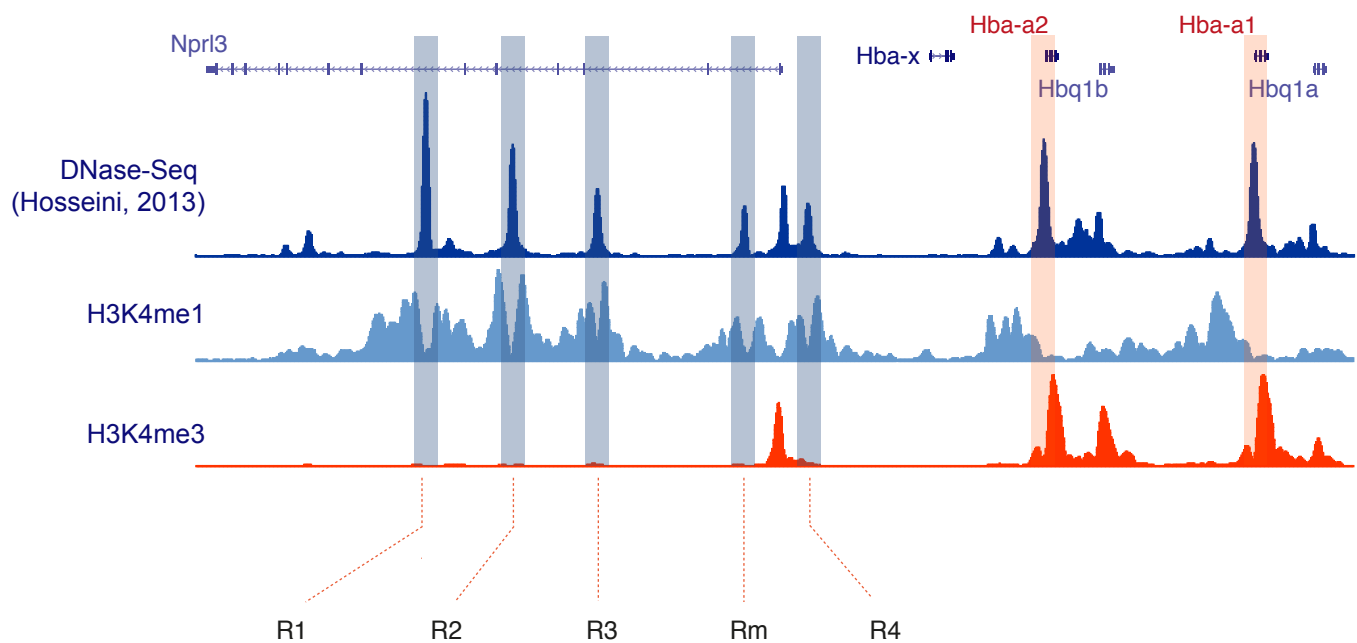


Figure 3. α -globin locus in the mouse. UCSC tracks showing the genomic region of the α -globin locus. DNase-Seq enrichment highlights the regulatory elements. ChIP-Seq of H3K4me1 (light blue) and H3K4me3 (red) annotate the 5 enhancers (R1 to R4 and Rm), as well as the α -globin promoters. RefSeq genes are shown on top.

(b) The regulatory elements are not equivalent

Recent studies in the lab have shown that the five regulatory elements are not equivalent. We demonstrated (Hay 2015, under review) using knockout models of the globin enhancers, quantification of nascent transcription from the globin promoter and analysis of the chromatin features, that even though the five enhancers appear the same based on genomic annotation (chromatin marks), their contribution to transcriptional regulation is different. The major drivers of α -globin expression are R1 and R2, while R3, R4 and Rm have only minor contribution and thus, undetermined functional roles.

Interestingly, the regulatory sequences underlying R1 and R2 are indicative of their important role, as these two are most enriched in erythroid TFs (Table1). This illustrates that even though enhancer activity depends on a permissive chromatin environment, the intrinsic features of the regulatory sequence plays an essential role in determining the contribution of an enhancer to transcriptional output.

The directly bound TFs are the most likely to be affected by variants in the sequence, and lead to defects of α -globin expression. We are interested in exploring the effect of SNPs on enhancer function (as discussed in the previous section) and we will do so by analysing the human α -globin cluster, where natural occurring variation can be interpreted in respect to the thalassemic phenotype.

(c) The α -globin cluster in the human

The human α -globin cluster is located at the telomeric end of chromosome 16 (16p13.3), and it includes an embryonic gene (ζ), two minor α -like genes [α D (also called μ) and θ], two α genes ($\alpha 2$ and $\alpha 1$), and two pseudogenes ($\psi\alpha 1$ and $\psi\zeta$)^{50,51}.

The syntenic region also contains, as described above, the four conserved regulatory elements as well as several widely-expressed, flanking genes⁵¹. Similar to the mouse model, the enhancers and promoters in the cluster are DNase hypersensitive in differentiated erythroid cells, and are marked by characteristic histone marks (Figure 4).

Table1: Binding of erythroid TFs at the regulatory elements of the human α -globin cluster

Element/Binding	NF-E2	GATA1	SCL	KLF	# TFs	Med1	H3K27Ac	H3K4me1	H3K4me3
R1	+	+	+	+	4	+++	+++	+++	
R2	+	+	+	+	4	+++	+++	+++	
R3		+	+	+	3	++	+++	+++	
R4		+	+	+	3	++	+++	+++	+
Rm	+	+			2	++	++	+++	
Promoters		+		+	2	+++	+++		+++

(i) Defects in human α -globin expression can lead to α -thalassemia

The integrity of the two pairs of α -globin genes ensures correct expression and mRNA production that result in the appropriate amount of α -globin chains.

Deletions or mutations of the adult α -globin genes that result in an imbalanced amount of α -globin chains can lead to α -thalassaemia or clinical disorders caused by α -thalassemia. The two types of α -thalassaemia are: (1) α^+ -thalassemia, in which only one of the pairs is affected ($-\alpha/\alpha\alpha$); this is also called the “silent carrier” state. (2) α^0 -thalassaemia, in which both genes are inactivated ($--/\alpha\alpha$); this leads to the α -thalassaemia trait⁵⁴.

Co-inheritance of both α^0 - and α^+ -thalassaemia results in three of the genes being inactivated ($-\alpha/--$) and leads to a severe form of anaemia, HbH disease. In HbH disease, excess β -globin chains form tetramers and precipitate in the blood, reducing red cell survival. The most severe genetic disorder is a result of homozygous inheritance of α^0 -thalassaemia, which inactivates all four α -globin genes ($--/--$). The deficiency in α -chains in the fetus leads to excess production of γ -chains (from the β -globin locus), which form γ -globin tetramers, or Hb Bart's. This causes Hb Bart's hydrops syndrome, a lethal condition if post-natal transfusions are not administered⁵⁴.

Deletions of the conserved regulatory elements, that leave the globin genes intact can also result in α -thalassemic phenotype, suggesting some of the elements are required for correct α -globin expression^{55,56}. The description of a small deletion that removes R1 and R2⁵⁵ demonstrated that one or both of these elements are necessary for the activation of the promoter.

Various experiments containing the human α -globin either in cell lines (Murine Erythroleukemia <MEL> hybrids containing the human chromosome 16)²⁸ or in mouse models (in particular the humanised mouse, with the native α -globin cluster replaced by the human α -globin cluster)⁵⁷ have led to the conclusion that R2 (HS-40) is the major regulatory element for α -globin expression. In these experiments, deletions of R2 result in failure to up-regulate of the human α -globin expression. These observations were supported by two Australian patient cases with deletions described by Phylipsen et al.,⁵⁶ who, out of the four regulatory elements, only lack R2 (with small proportion of Nprl3 coding sequence). However, the surprising case of a Portuguese patient who has the R2 mutation in homozygosity, has confused the field as such a mutation was not expected to be viable with life. The patient has HbH disease, but is otherwise well. Numerous questions have emerged from the description of this last case, the Portuguese patient that lead to questioning the R2 in regulating human α -globin. Though, it may be that R2/HS-40 is not critical for expression, it is clear that the element is necessary for correct α -globin expression.

This role of R2/HS-40 is also supported by population studies. A recent red blood cell GWAS⁵⁸ has revealed significant associations for common SNPs in HS-40 with

red blood cell phenotypes. These variants have an effect on red blood cell parameters, also used to diagnose anaemia: mean cell volume (MCV), mean cell haemoglobin concentration (MCHC) and red blood cell count (RBC). Thus, even if R2 may not be critical (as showcased by only one patient for now), alterations as small as point mutation within R2/HS-40 can lead to susceptibility to anaemia at a population scale⁵⁸.

The nature of such SNPs in terms of what sequences they affect and how they impact in TF binding sites are of interest to us as we can investigate variation in the major regulatory element of α -globin, R2 (HS-40). This enhancer sequence has been well characterised, alterations can be interpreted in light of thalassaemic trait, and its function has been described in both human and murine models of erythropoiesis.

1.06 Erythropoiesis, model system for molecular genetics

Studies of transcriptional regulation for haemoglobin synthesis and defects in globin gene regulation leading to hemoglobinopathies have set the basis of molecular genetics. The first disease understood at a molecular genetic level was a red blood cell disease- sickle cell anaemia. Pauling et al., published in 1949 a study linking a defective molecule (an abnormal form of haemoglobin) to sickle cell disease ⁵⁹. At the dawn of molecular genetics, the globins (especially β -globin) provided the main model for gene regulation studies. Characterising the globin genes has drawn interest not only from a clinical perspective, as hemoglobinopathies are the most common monogenic disorder ⁶⁰, but also from a basic science perspective, as they represent excellent models for transcriptional regulation. Understanding the molecular basis of the globin regulation has furthered our understanding of the mechanism by which the thalassemias arise. They also elucidated the major principles that govern gene transcription and regulation.

Erythropoiesis offers a convenient and accessible model system, as the stages of differentiation are well characterised and both primary cells and derived cell lines for different stages are available for both mouse and human. Moreover, we can study both aspects of development by comparing primitive and definitive erythropoiesis, and aspects of differentiation, following progenitor cells from commitment to differentiation (for example, in definitive erythropoiesis discussed below).

(a) Transcriptional regulation of in erythropoiesis

Lineage specific differences in gene expression are driven by binding of TFs to regulatory sequences. The binding of erythroid-affiliated TFs has been characterised, initially with a focus on globin genes regulation and more recently, globally in erythroid tissues. Some of these TFs are involved in the regulation of erythropoiesis in both development (GATA1, SCL/TAL1, KLF1, LMO2, LDB1) and differentiation. We summarise briefly three of the main players in erythropoiesis. A list of important erythroid TFs can be seen in Table 2 below.

GATA1 is a C2H2 zinc finger TF that binds promoters and enhancers of most of the erythroid specific genes⁶¹. This master regulator of erythropoiesis is necessary for the function of numerous red cell specific genes as illustrated by transcriptome analysis of the G1-ER cell line⁶². GATA1, is involved in transcriptional regulation of both primitive and definitive cells, its disruption leading to maturational arrest at the proerythroblast stages^{63,64}. Regulation by GATA1 is achieved by directly binding of DNA at the WGATAR motif^{65,66}.

KLF1 (or eKLF) like GATA1, is critical for erythropoiesis⁶¹. KLF1 has been shown to play a role in (1) erythroid lineage commitment, at the megakaryocyte-erythroid progenitor stage; (2) global expansion of erythroid gene expression in both primitive and definitive cells, with direct involvement in globin switching; (3) in terminal maturation by controlling exit from the cell cycle⁶⁷. KLF1 binds both promoters and enhancers (including the globin locus-control regions) at different stages of erythropoiesis, as shown by ChIP-Seq⁶⁸. This study refined the preferred CACC-binding motif to CC[C/A]CGCCC.

SCL/TAL1 is a bHLH TF involved in transcriptional regulation of different stages of hematopoiesis such as early haematopoiesis in which it dimerises with E proteins (E2A), but also in later committed, committed stages such as the erythroid lineage. SCL/TAL1 has been most extensively studied in erythropoiesis where it is involved in the early development of both primitive and definitive erythropoiesis, as demonstrated by various knockout experiments^{69,70}. SCL/TAL1 regulates expression of genes in most red-cell specific processes both through direct binding of DNA (majority), but also by being tethered at regulatory sequences⁷¹. Sequences preferentially bound by SCL *in vivo* are the E-box [CANNTG] with the variation CAGCTG or CAGGTG⁷², as well as composite motif containing a GATA site and half E-box (CTG) separated by 9bp⁷¹.

The relevant aspect for this thesis is the direct binding of these TFs to preferred DNA motifs, in HS regions, and if/how this binding changes through dynamical processes.

Table 2: TFs involved in erythropoiesis (Adapted from Higgs et al., 2008).

TF	Class	Binding motif	Transcriptional Regulators			Key References
			Ubiquitous	Primitive	Definitive	
Gata1	Zinc Finger (C2C2)	WGATAR		x	x	(Cantor and Orkin, 2002; Orkin 2000)
Gata2	Zinc Finger (C2C2)	WGATAR			x	(Cantor and Orkin, 2002; Orkin 2000)
Scl/Tal1	bHLH	CAGGTG (E-box) CAGCTG (E-box)		x	x	(Cantor and Orkin, 2002; Wadman et al., 1997, Kassouf et al., 2010)
E2A	bHLH	CAGGTG (E-box)	x		x	(Andrews, 1998; Motohashi et al.,2002)
p18 (NF-E2)	b-Zip	TGCTGACTCAT	x		x	(Bieker 2001, Tallack, 2010)
p45 (NF-E2)	b-Zip	TGCTGACTCAT			x	
Klf1	Zinc Finger (C2H2)	CC[C/A]CGCCC		x	x	(Suske, 1999; Van Loo et al.,2003)
Sp1	Zinc Finger (C2H2)	GGGGCGGGG	x		x	(Woo et al.,2005)
Sp3	Zinc Finger (C2H2)	GGGGCGGGG	x		x	
ZBP-89	Zinc Finger (C2H2)	<<G-rich box>>	x		x	(Kabe et al, 2005; Mantovani 1999)
NF-Y	Heterodimeric CCAAT factor	CCAAT	x		x	(Yi et al., 2006)
Sox6	HMG			x	x	(Sankaran et al., 2009, Xu et al.,2010)
Bcl11A	Zinc Finger (C2H2)	GGCCGG		x	x	(Cantor and Orkin, 2002; Orkin 2000)
FOG-1	Multi-type Zinc Finger	Co-factor			x	(Nam and Rabbits, 2006)
Lmo2	LIM domain	Co-factor		x	x	(Visvader et al., 1997)
Ldb-1	LIM domain	Co-factor	Widely expressed	x	x	

* Abbreviations: b-Zip: Basic Leucine Zipper factors; bHLH: basic Helox-Loop-Helix factors

(b) Development of red blood cells

Red blood cells, the most abundant cell type in the body, are of two different kinds based on the developmental stage they emerge at. Primitive cells are transient early in the embryo and circulate as nucleated, larger cells which eventually enucleate. Definitive red blood cells, found in circulation smaller and enucleated, at fetal and adult stages⁷³.

(i) Primitive erythropoiesis

The first blood cells in the mouse emerge in “blood islands” in the yolk sack, and have been observed as a transient wave between E7.25 and E9.0. This unique population, termed primitive erythroid colony-forming cells (EryP-CFC)⁷⁴ is followed by a semi-synchronous wave of maturing nucleated primitive precursors between E9.5-E12.5. In a similar manner to the definitive erythroid compartment, these precursors mature in circulation by symmetric cell division, accumulation of hemoglobin, decrease in cell size, nuclear pyknosis and decrease in RNA content⁷³. By E12.5, the population matures into orthochromatic erythroblasts and cell divisions stop. And despite the long-held belief, the primitive cells also enucleate in the final stages of their maturation (E12.5 - E16.5)⁷⁵.

(ii) Definitive erythropoiesis

Definitive erythropoiesis occurs in two waves. First wave is transient and occurs in the yolk sack, from where definitive erythroid progenitors expand to the fetal liver.

The second wave emerges from major arterial vessels, expands in the fetal liver until approximately E16.5 and then transitions to the bone marrow. The transition to the long-term niche, the bone marrow begins by E17.5 and is completed after birth⁷³.

The earliest definitive progenitors, committed to the erythroid lineage are the burst-forming unit erythroid (BFU-E) and colony-forming unit erythroid (CFU-E). As suggested by their name, these progenitors were isolated by their ability to form erythroid colonies in semi-solid media. In terms of differentiation, BFU-Es are the more immature population as judged by differentiation *in vitro*^{76,77}. CFU-Es most closely correspond *in vivo* to proerythroblasts, the earliest recognisable erythroid progenitor in fetal liver and bone-marrow. Once formed these cells will progress to definitive erythroid precursors and finally red blood cells. The late-stage precursors are defined based on morphological characteristics of the cell volume, nuclear condensation and accumulation of haemoglobin and follow the progression from: pronormoblasts (proerythroblasts), normoblasts (basophilic, polychromatophilic), late normoblasts (orthochromatic erythroblasts). The last stage of enucleated cells corresponds to reticulocytes which will lead to the formation of the biconcave, hemoglobinised red blood cells (RBC)⁷⁸. The expression of α -globin increases with progressive erythroid maturation, and the process of its distal regulation can be followed from BFU-E/CFU-E stage when the promoter is expressed at basal levels, to normoblast/latenormoblasts mix when α -globin expression is enhanced.

(iii) The switch from primitive to definitive erythropoiesis

Aside from the different morphological features, one of the main characteristic of primitive and definitive erythroid cells is that they express different globin genes from both α - and β -clusters. At the α -globin cluster, the primitive cells express embryonic ζ - and α -globin genes, while the definitive cells express only α -globin genes for both mouse and human. At the β -globin cluster, the expression differs between mouse and human. In primitive cells the genes expressed are the embryonic ϵ y- and β H1-globin in the mouse, and ϵ - and γ -globin genes in the human; in definitive, “adult” β 1- and β 2-globin genes are expressed in the mouse, and “fetal” γ - and “adult” β -globin genes are expressed in the human⁷⁹. The mouse, like most species, have only one change from embryonic to definitive erythropoiesis, whereas the human has an additional one, to the γ -genes in the fetal period. The expression of fetal γ -globin is a relatively recent event that occurred in primates evolution 35 to 55 million year ago⁸⁰ (Figure 5).

The transition from embryonic to fetal and to adult globin has been termed “switching” and has been extensively studied due to its therapeutic potential but also aspects of transcriptional regulation. The process of “switching” is an interesting transcriptional regulation problem, as the same regulatory regions (α -/ β -globin LCRs) are involved in regulating different promoters at different stages. How this regulation is achieved is yet to be determined, but the binding of TFs to the enhancers in primitive and definitive cells is most likely one of the central aspects.

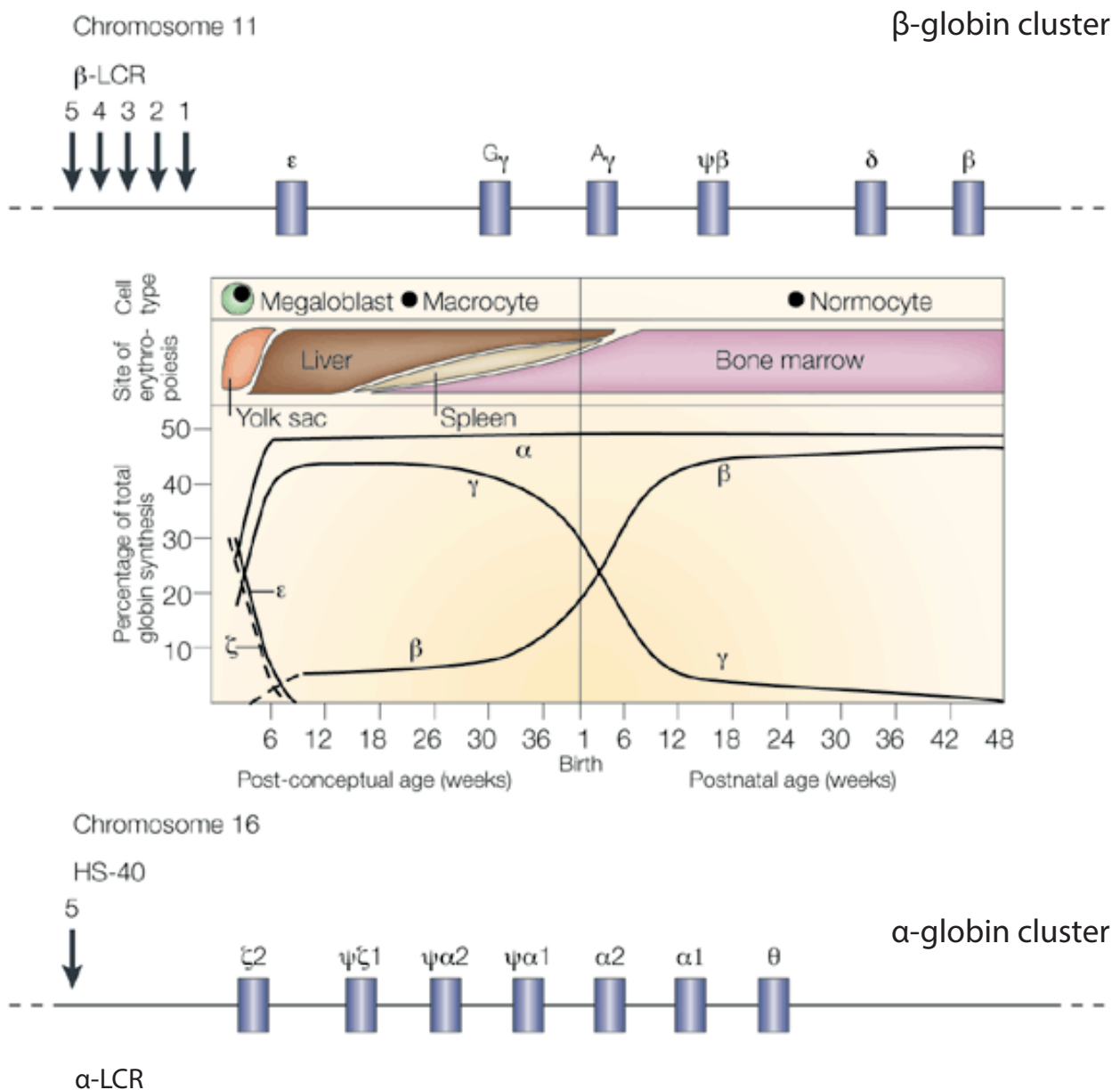


Figure 5. The developmental switch in human, from embryonic to fetal and adult globin genes. The α -globin gene cluster on chromosome 16 and the β -globin gene cluster on chromosome 11. Vertical arrows indicate the location of DNaseI hypersensitive sites that are thought to be involved in globin gene regulation. The products of the G_γ - and A_γ - genes are γ -chains with either glycine (G_γ) or alanine (A_γ) at position 136. The insert shows the sequential activation of the embryonic, fetal and adult globins. A megaloblast is a large red- cell precursor, a macrocyte is a large red cell and a normocyte is a normal-sized red cell. (LCR, locus control region). This figure was adapted from (Weatherall, 2001).

1.07 Differentiation of definitive red blood cells

(a) Transcription factor binding at the *cis*-regulatory elements through erythropoiesis

Previous work from our lab has addressed, in primary cells from different sources and murine erythroleukemia cells (MEL) the sequence of events in mouse erythropoiesis, leading from the silenced locus, in early, uncommitted progenitors, to the fully activated state in terminally differentiated cells. At ES stage, the α -globin cluster is not accessible, and none of the regulatory elements are detectable. In multipotent haemopoietic cells, the cluster become accessible to a certain extent with only some of the elements in the cluster (R2 and R4) in an open chromatin conformation. Erythroid specific TFs (GATA2, NF-E2, SCL complex) bind the accessible regulatory regions, and the histones become acetylated. At some point before the cells reach the fully committed erythroid progenitor stage, GATA2 is replaced by GATA1 as a component of GATA1/SCL pentameric complex, the rest of the hypersensitive sites are formed at the enhancers (except for R3) and the promoters, and histone acetylation extends throughout the region. The entire α -globin cluster, with the regulatory regions is now prepared for transcription: the chromatin is marked for transcriptional activation but there is no globin mRNA being produced, until terminal differentiation, when PolII is recruited to the α -globin promoters and transcription starts^{53,81,82}.

As the cells progress through these stages of erythropoiesis, the process of TF binding to the regulatory sequences within the hypersensitive sites appears to be very dynamic including GATA2 to GATA1 switch, recruitment of more TFs (KLF).

However, it is unclear from ChIP-qPCR and ChIP-Seq experiments, which of these elements are directly bound to DNA, and which are involved only in protein protein interactions. Are all the TF binding sites within the HSs occupied by TFs, or only some? If not all, which ones? Are the conserved sites occupied or the rather the species specific sites? These questions cannot be addressed only by ChIP-Seq but rather by an integrative approach using ChIP-Seq and high resolution analysis of the TF binding sites footprints.

Aim

The aim of the current thesis is to analyse the changes in TF direct binding at regulatory regions in development, differentiation and disease.

To this end we study the protein-DNA interaction at regulatory regions (enhancers, promoters) from the DNA perspective. Within regulatory regions, or hypersensitive (HS) regions, proteins bound to DNA will protect the underlying site from DNaseI digestion. If sequenced at sufficient depth, the frequency of DNaseI cutting can be used to identify the regions where a TF is directly bound to DNA. We have developed novel methods to interpret TF binding at regulatory regions and apply these to erythroid cells in different contexts.

We analyse direct TF binding at high resolution employing two approaches, both based on open chromatin data: (1) “digital footprinting”, which shows a footprints corresponding to a specific genomic region; (2) a novel approach we call “average footprinting” which identifies the cell-specific DNA footprint for any sequence of interest by piling-up and averaging all the digital footprints of that sequence.

In **Results Chapter 1** we present the development and setup of the open-chromatin assays (DNase-Seq and ATAC-Seq) necessary for the isolation and identification of HS regions and their analysis at low and high resolution, from different cell types. The performance of the two methods both at the model locus and genome-wide is compared, with advantages and limitations.

Results **Chapters 2 & 3** describe the analysis of the differences in enrichment and binding pattern at HS regions, in development or differentiation respectively, at the α -globin locus and genome-wide. The findings are reported following a similar structure in both chapter

Firstly, we visualise and quantitate the enrichment of open-chromatin in HS regions at the α -globin locus and relate the observations to the transcriptional state of the α -globin gene (basal/expressed/not expressed). Secondly, we assess direct binding of TFs at high resolution by analysing the DNase/ATAC digital footprints

At a genome-wide level we assess cell specific binding of TFs by comparing HS regions. The statistically significant regions that are cell type specific are annotated (enhancers, promoters) and characterised in terms of their location in the genome and relation to known erythroid genes.

In **Results Chapter 4** we present a novel approach for assessing direct binding to DNA by “average footprints”. The “average footprints” measure frequency of binding of a sequence and thus reflect usage of that sequence. We apply this approach to identify the regulatory regions and motifs used preferentially in a cell type.

In **Results Chapter 5** we investigate the effect of SNPs in regulatory regions on gene expression. We assess in silico the impact of sequence variation on direct binding of TFs. To this end, we focus on the α -globin locus in the human and a patient with the thalassemic trait. This patient with unexplained anaemia has a variant in the main enhancer and none of the known α - or β - deletions. We present an adaptation of the “average footprints” approach, to predict potential damage caused by this variant. Finally, we present preliminary work to functional validate the impact of this variant on TF binding.

II. Materials and Methods

1.01 Isolation of cell populations

Primary cells were isolated from C57Bl/6 mice spleen or fetal livers. All animals were housed and handled according to institutional and national guidelines for humane animal care.

Cell populations isolated were characterised in terms of immunophenotyping (FACS), morphology (MGG staining of cytopins) and expression of α -globin (Real Time PCR) prior to genomic analysis.

(a) Mouse erythroid cells

(i) Murine erythroleukemia (MEL) cells

MEL cell line 585^{28,83,84} were grown in RPMI 1640 medium (Gibco, Life Technologies) supplemented with 10% foetal calf serum (FCS), 50 U/ml Penicillin, 50 μ g/ml Streptomycin, and 2mM L-Glutamate.

(ii) Splenic erythroblasts

Splenic erythroblasts were isolated from phenylhydrazine treated C57Bl/6 mice. Phenylhydrazine injections were performed as described in⁸⁵: 0.04mg of l-acetyl-2-phenylhydrazine(Sigma) dissolved in Hank's balanced salt solution per gram of mouse weight (1mg per average 25g mouse) was injected intraperitoneally three times over a period of two days. Spleens were harvested on day 6.

Spleens were disaggregated using a scalpel, filtered through a 70nm filter (CellTrics) and stained using a PE anti-mouse Ter-119 (Ly-76, BD Pharmingen) at

10^7 cells/100 μ l staining buffer (PBS/ 2mM EDTA/ 0.5% BSA) for 20min on ice. Cells were then washed and incubated in the same buffer with goat anti-rat MACS beads (Miltenyi Biotec) and rotated for 15min at 4°C. Selection of the Ter119+ population was achieved on LS MACS columns (Miltenyi Biotec) according to manufacturer's instructions. Purity of the population was assessed by MGG staining and flow cytometry.

(iii) Embryonic fetal liver progenitors

E14.5 fetal liver cells were disaggregated by vigorous pipetting, filtered through a 70 μ m filter (CellTrics) and stained using a PE anti-mouse Ter-119 (Ly-76, BD Pharmingen) at 10^7 cells/100 μ l staining buffer (PBS/ 2mM EDTA/ 0.5% BSA) for 20min on ice. Cells were then washed and incubated in the same buffer with goat anti-rat MACS beads (Miltenyi Biotec) and rotated for 15min at 4°C. Separation of the desired progenitor population was achieved using a succession of antibody staining and magnetic columns (Miltenyi Biotec) as depicted in Figure 6 below.

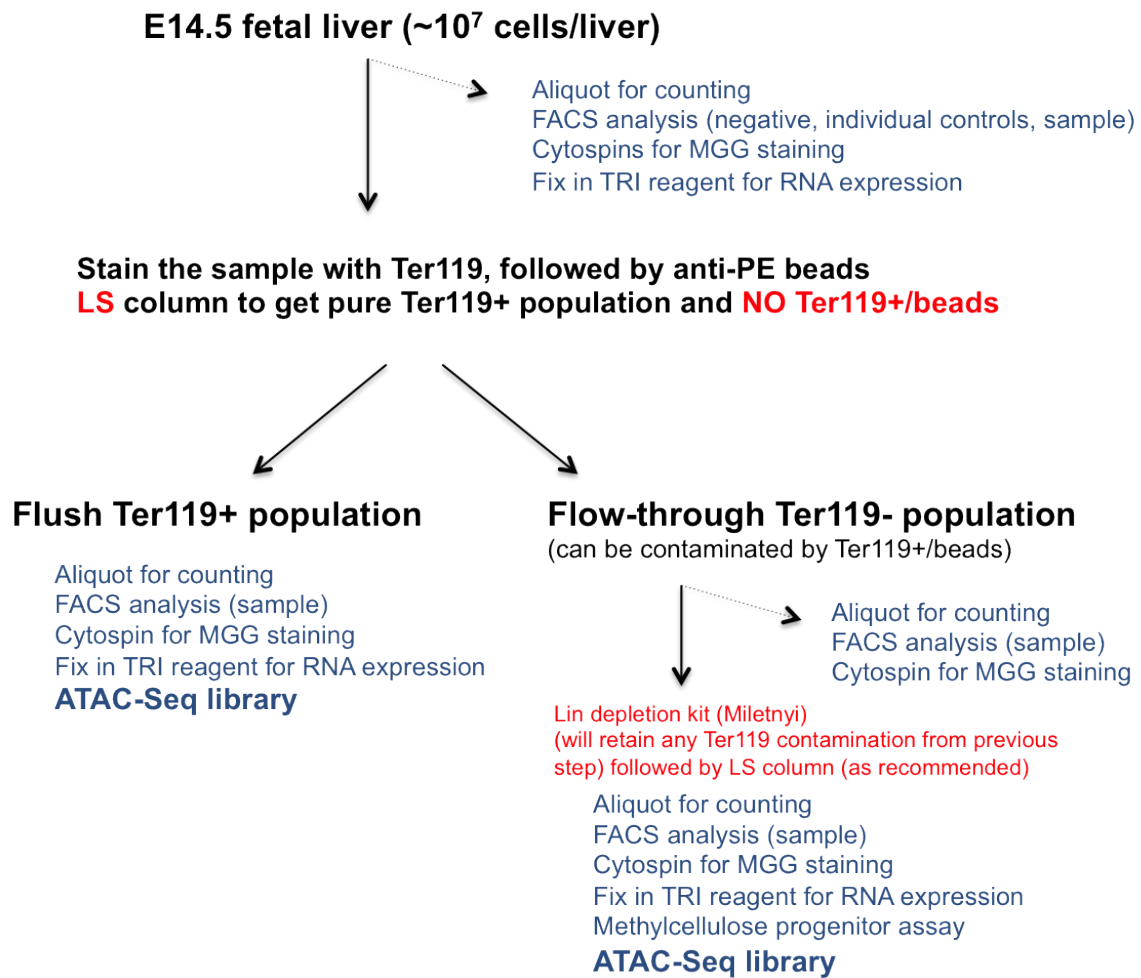


Figure 6. Strategy for isolating the early progenitors and differentiated cells. Pure populations were isolated from E14.5 fetal liver, based on a succession of antibody labellings and a magnetic column separations. The Ter119-, Lin depleted population corresponds to the early progenitor cell and the Ter119+ population corresponds to differentiated cell.

(iv) ES cells

In vitro differentiated E14 WT ES cells, harvested on day 3, were used for open chromatin assays as a non-erythroid control.

(b) Human erythroid cells

(v) Primary erythroid culture

Human erythroid progenitors were expanded and differentiated in vitro by J Sloane-Stanley following the Fibach protocol⁸⁶. Two phase human erythroid progenitors cultures were prepared from approximately 50ml of peripheral blood. In summary, mononuclear cells are isolated by centrifugation of peripheral blood on a histopaque gradient. In phase I, the isolated cells are cultured in phase I medium: DMEM supplemented with L-Glutamine 2mM, 10% FCS, 1mg/ml cyclosporin A and 10% conditioned medium from cultures of 5637 bladder carcinoma cell lines (ATCC) at 37°C, with 5% CO₂. In phase II, the non-adherent cells are harvested and re-suspended in phase II medium: DMEM supplemented with 30% FCS, 1U/ml human recombinant erythropoietin, 1µl per ml culture of beta-mercaptoethanol 1x10⁻⁵ M, dexamethasone 1x10⁻⁶ M final concentration, holo-transferrin 0.3ug/ml final concentration, recombinant human stem cell factor 10ng/ml and 1% final concentration of BSA.

For analysis, the samples were collected at the intermediate Fibach stage (day 8-9 of phase II).

1.02 Quantitative Real-Time PCR

Total RNA was isolated from 2×10^5 cells (Ter119+ population and ckit+ population isolated as previously described) using the TRI Reagent (Sigma) according to manufacturer's instructions. In order to avoid contamination with genomic DNA during the procedure, DNase treatment was carried out on the purification column. The resulting RNA was reverse-transcribed (RT) using the Superscript III First Strand Synthesis (Life Technologies) according to manufacturer's instructions. Primers designed and tested by Dr Bryony Graham over the intron-exon junction of the α -globin promoter (Table 3) were used for Real-Time PCR experiments. Reactions were setup containing 2 μ l of sample or noRT control and 23 μ l of a mix containing 200nM primer in SYBR Green PCR Master Mix (Applied Biosystems). Every sample was analysed in duplicate and run on the StepOne Real-Time PCR System (Applied Biosystems) using StepOne software. Cycling conditions were 50°C for 2min and 95°C for 10 min followed by 40 cycles of 95°C for 15 seconds and 60°C for 1 min. Quantification of the relative abundance of detected sequences was calculated using the $\Delta\Delta C(t)$ method where $C(t)$ is the number of cycles above which fluorescence of PCR products can be detected above background. If the amplification efficiency of each cycle is close to 2, the relative abundance of the gene of interest ($C(t)_{\text{target}}$) compared to a housekeeping control ($C(t)_{\text{reference}}$) can be calculated as $2^{-\Delta C(t)_{\text{target}}} / 2^{-\Delta C(t)_{\text{reference}}}$ ⁸⁷.

The $C(t)$ value for every sample was averaged between duplicates for a specific primer set. Relative gene expression ratios were calculated to ribosomal RNA 18S.

Table 3: Primer sequences of primers used for α -globin expression analysis

Gene	Protein	Forward	Reverse	Size [bp]
Hba-a1	Haemaglobin alpha 1/2	GTGTGGATCCCGTCAACTTC	CCACTATGTTCCCTGCCTTG	108
		CAAGGCAGGGAACATAGTGG	GGCAAGGAATTTGTCCAGAG	162
Hbb-b1	Haemaglobin subunit beta-1	GCACCTGACTGATGCTGAGA	TCTCCAAGCACCCAACTTCT	140
		ATGGCCTGAATCACTTGGAC	AGAATAGCCAGGGGAAGGAA	141
Rm18s	18s ribosomal RNA	GTAACCCGTTGAACCCATT	CCATCCAATCGGTAGTAGCG	151

1.03 Cellular staining

All cellular staining procedures were performed on material derived from at least three independent experiments/animal tissues for all genotypes tested.

(a) May-Grünwald Giemsa (MGG) staining

May-Grünwald Giemsa (Sigma) is an acid-base staining used on haematological samples for morphological assessment. Cytospins are prepared on slides with approximately 10^5 in 200µl medium/ PBS with 10% FCS and cytospun onto slides. After drying the slides can be stained with MGG according to manufacturer's instructions, mounted and examined microscopically.

(b) Image analysis

Images of cytopins were taken using a Leica microscope, objective 20x magnification.

(c) Immunophenotyping in suspension for FACS analysis

Cells isolated from embryonic fetal livers were immunophenotyped directly then analysed by FACS using Cyan Flow Cytometer (DAKO cytometry) equipped with the software summit 4.2 (DAKO cytometry). Staining was performed in round-bottom 96-well plates. Approximately 10^6 cells were resuspended in 100µl of staining buffer (PBS with 10% FCS) containing the following antibodies: PE anti-mouse Ter-119 (Ly-76), FITC anti-mouse CD71 (C2), APC anti-mouse c-kit

(CD117)(2B8) (all from BD Pharmingen). Single staining with biotin-conjugated Gr-1 (Ly-6C) or biotin-conjugated CD11b (Mac-1)(BD Pharmingen) was followed by Streptavidin-PE (BD Pharmingen) staining.

Spleen cells were stained with PE anti-mouse Ter-119 (Ly-76). Controls for each isotope, as well as negative controls were prepared for calibration of the flow cytometer.

The fluorescence emission for FITC, PE, PE-Cy5 (TC), and APC was detected through 530/30, 580/30, 670/30, 670/40nm band pass filters respectively and 740nm long pass filter for APCCy7. Data was acquired with Summit Software provided by DAKO Cytomation and analysed with FlowJo. In cases where overlap between fluorochromes was observed, compensation was adjusted accordingly prior to acquisition. Relevant isotypes were checked for all antibodies used in this assay.

1.04 Embryonic fetal liver progenitors assay

2×10^5 E14.5 fetal liver cells were plated in 0.8% methylcellulose containing serum, transferrin, and insulin (M3234 Stem Cell Technology) and was supplemented with a cocktail of growth factors consisting of Tpo (1% of conditioned cell culture supernatant), 0.02 or 2U/ml Epo (Roche), 50ng/ml KL, 10ng/ml IL-3, 5ng/ml IL-11, 2ng/ml IL-6 (all Peprotech), 10ng/ml Flt-3-ligand (R&DSYSTEMS).

CFU-E were scored on day 3, BFU-E, megakaryocytic and other myeloid colonies from day 7 onwards, based on morphological characteristics subsequently confirmed by MGG staining and immunophenotyping and FACS analysis.

1.05 Open-chromatin assays

(a) DNaseI assay

DNaseI assay was performed with a starting material of $30\text{-}50 \times 10^6$ cells as previously described³⁴. In summary, cell nuclei were isolated by incubation with in high salt lysis buffer (50mM KCl, 10mM MgSO₄, 3mM DTT, 5mM Hepes, 0.05% Igepal CA 630), washed with RSB buffer (10mM NaCl, 10mM Tris pH7.4, 3mM MgCl₂.6H₂O) and prepared for digestion with different concentrations of DNaseI enzyme (Roche).

(vi) DNaseI digestion

DNase I digestion was performed with different concentrations of DNaseI recombinant, RNase free (Roche) in 1ml reaction volume at 37°C until the median sample reached water-like viscosity. Digestion was stopped by placing the samples on ice. Reactions were incubated overnight at 37°C in lysis buffer and proteinase K. The next day, samples were incubated for 30min with RNaseA and DNA isolated by phenol/chloroform extraction and ethanol- precipitation.

(vii) DNaseI hypersensitivity at a single HS region

DNaseI digestion was assessed by Southern Blot as previously described in ²⁸.

The DHS sites are identified by Southern Blotting, using indirect end-labelling. This method consists of using a probe that binds to a heterogeneous mixture of digestion products. The probe was designed to hybridise to the digest fragments containing a HS. If the fragment was not digested by DNaseI, a band of the size of the entire

digest fragment would be visible on the Southern blot. If the HS was exposed to DNaseI digestion, the fragment would be cut and smaller bands of known different sizes would appear on the blot.

Selection of fragments for library preparation based on degree of digestion and size

Degree of digestion was assessed using the TapeStation Genomic DNA kit (Agilent) and the DNaseI digested fractions to be used for library preparation were selected based on the reduction in concentration of the fraction containing very large genomic DNA fragments (>40Kb).

(viii) Library preparation

The library preparation was performed using the NEBNext DNA library preparation (E6000S) with adaptations for low input DNA and size selection. Adapter stocks were diluted 1:10 and primer stocks were diluted 1:10. For size selection, libraries were run a 2% agarose gel (Biorad), stained with Ethidium Bromide (Sigma) and the required size (145bp-450bp) excised. DNA was extracted using a gel extraction kit (Zymo).

(ix) Library quality control (QC)

To ensure optimal yield during sequencing, the libraries must be quantitated. Amounts loaded for a sequencing run on Illumina platforms (MiSeq, NextSeq, HiSeq) were usually 4nM total library, in a volume of 5ul.

For quantitation, library concentration and fragment size must be calculated in order to calculate molarity. The molarity can be finally checked by running a custom library qPCR.

For the library concentration, a broad range kit for the Qubit spectrophotometer (Invitrogen) was used and measurement was performed following manufacturer's instructions. Library fragment size was detected using a D1000 tape for the Tapestation (Agilent) following manufacturer's instructions.

Based on this two measurements, the molarity of the library was calculated and 4nM stock was prepared. The 4nM stock was further checked using the universal library quantification kit from Kapa Biosystems that uses primers designed to sequencing adaptors to determine the amount of library that will be sequenced. The concentration can be calculated in relation to DNA standards of known concentrations. In case of discrepancies, the final amount loaded on the machine was adjusted accordingly to have a final concentration of 4nM according to the qPCR quantification.

(b) Assay for transposase-accessible chromatin sequencing (ATAC-Seq)

Assay for transposition of active chromatin sequencing (Figure 7B) was performed as previously published ⁴⁴. $6-8 \times 10^4$ cells were used per biological replicate. Cells were lysed and nuclei were isolated prior to transposition with Tn5 transposase (Nextera, Illumina) for 30min at 37°C. DNA was purified using a MinElute kit (Qiagen). Libraries were amplified and barcoded using the NEBNext 2xMastermix (NEB) and the custom primers as published in Buenrostro et al., 2013.

ATAC-Seq libraries profiles were visualized using D1000 tape on the Tapestation (Agilent). Due to the broad range of DNA fragment sizes found in these libraries, quantitation with the Qubit for DNA concentration was found to be highly variable and was omitted. The libraries were quantified using the universal library quantification kit (KAPA Biosystems).

1.06 Naked DNA controls for DNase-Seq and ATAC-Seq

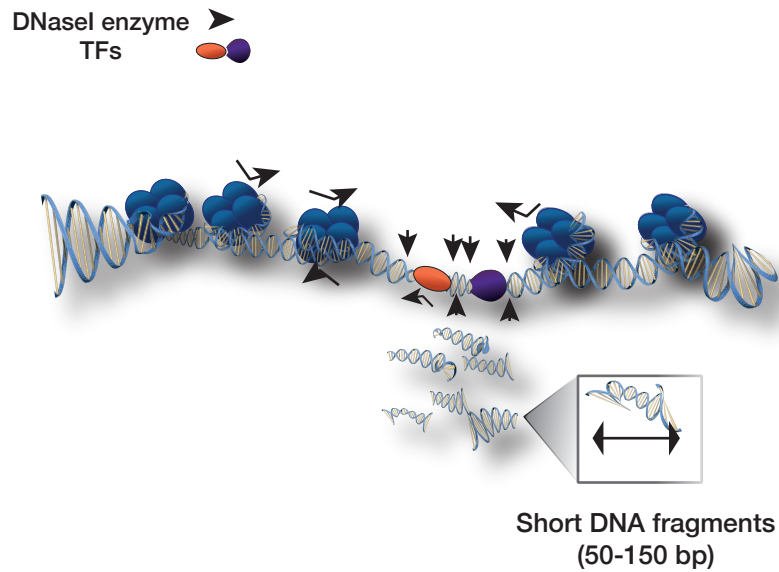
Naked DNA controls were generated for both DNase-Seq and ATAC-Seq. For DNase-Seq control, digests of 2ug genomic DNA were incubated with 20, 30 and 40U of DNaseI. Libraries were prepared as described above.

For ATAC-Seq 100ng of genomic DNA was incubated with the Tn5 transposase and the library was prepared as described above.

1.07 Sequencing

Samples were sequenced using: 75bp paired end reads (MiSeq), 40bp paired end reads (NextSeq) and 50bp paired-end (HiSeq).

A)



B)

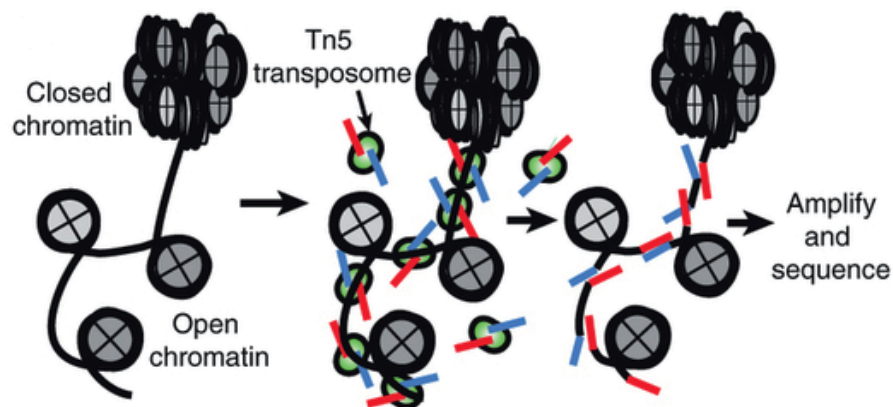


Figure 7. Illustration of the two open chromatin assays used in the study. A) DNaseI assay involves incubating live nuclei with the highly processive DNaseI endonuclease (black arrows) which will cut open chromatin preferentially. This process releases short hypersensitive fragments that can be processed for subsequent Southern blotting or sequencing. B) Assay for transposition of active chromatin (ATAC) involves incubating live nuclei with hyperactive form of Tn5. This process generates fragments with sequencing adaptors already ligated and can be sequenced. The transposition figure was adapted from (Buenrostro et al.,2013)

1.08 Southern Blot

Standard protocols were followed⁸⁸. Briefly, the digested samples from the DNaseI digestion assay were loaded on 1% agarose gels, one lane per gel (10 samples). After electrophoresis of the digested samples, the gels were denatured for 1hr (in 1M NaOH), neutralised (in neutralisation solution: 1M Tris, 3M NaCl, pH 7.4) for 1hr. DNA was transferred onto Zeta-probe membranes (Biorad) by upward capillary transfer. The membranes were baked for 2hrs in an 80°C oven. Location of the probes used is shown in Results Chapter 1 (Figure 9). They were generated as PCR products, labelled using the Megaprime Labelling Kit (Amersham, GE Healthcare). Hybridisation was carried out in a formamide based buffer at 65°C.

1.09 Computational analysis

(a) Genome Alignment

Paired-end reads for DNase-Seq, ATAC-Seq and genomic DNA control samples were processed using our in-house customised DNase and CHIP pipeline (Telenius, unpublished): samples were aligned to the appropriate genome build (mm9 for mouse data and hg18 for human) using bowtie (version 1.0.0, <http://bowtie-bio.sourceforge.net/index.shtml>)⁸⁹. To prevent the exclusion of the duplicated globin genes bowtie was run with the `-m` reporting option set to 2 to allow reads to map twice to the genome. To exclude over-amplified products from these data sets, reads that map to the exact same genomic position were collapsed into a single representative read. Ploidy regions which represent regions in the genome which

strongly overreact in high-throughput sequencing experiments due to large copy number differences between the real genome and the genome build and normalize poorly were excluded at this stage from downstream analysis. For the human data we used an Encode generated track (Duke Mappability Exclusion Track) while for the mouse data we generated a similar set of regions in-house.

(b) Genome-wide tracks

Genome-wide tracks of the DNase-Seq and ATAC-Seq were produced using the in-house perl tool sam2bigwig.pl, which produces a track of read density over a set window size and increment of movement across the genome. For the hypersensitive sites a window of 300bp and an increment of 30 bps was used. Bowtie alignments were thus converted to genome wide density tracks (BigWig) and the raw output was displayed in UCSC Genome Browser⁹⁰.

(c) Normalisation

Data sets were normalised for same number of reads, by random subsampling to the smallest data set using the in-house Perl tool resample.pl, version 2. These normalised data sets were then used for the differential analysis.

For visualisation in UCSC, the signal in the data sets was further normalised per million reads using the script makeUCSCfile.pl from the Homer package. The output is a normalised bedgraph file that can be visualised in UCSC Genome Browser.

Peak Calling

Peak detection for the DNase-Seq and ATAC-Seq in erythroid mouse and human fetal cells was performed with the MACS2 algorithm⁹¹. The test data sets were the

data sets normalised for read number, with naked DNA control tracks matched for the read number.

(d) Subsampling

Random subsampling of control data set to a lower number of reads was performed using the Samtools package. The option 'view' was used to subsample to 30% and 40% of the control bam file. The resulting bam files were then converted to sam files and bigwigs for visualisation.

The annotation of the genomic elements was performed based on the data from Ter119+ cells, combining two criteria: difference of chromatin marks H3K4me1 (associated with enhancers) and H3K4me3 (associated with promoters) enrichment from ChIP-Seq data previously generated in the lab, combined with genomic location in terms of distance to TSS, intergenic, intronic or exonic - a combination of RefSeq and UCSC annotations. Elements that could not be annotated using any of these criteria were classified as "not annotated".

For the heatmaps in the differentiation chapter, publicly available data sets⁹² comprising of H3K4me1, H3K4me2, H3K4me3, H3K27ac were used for the annotation of functional elements.

(e) Differential analysis

Differential analysis of HS regions was performed on three biological replicates using the R package DiffBind^{93,94}, with the analysis methods DESeq2. Peaks with less than 10 reads per million per killable were filtered out prior to quantification.

Normalisation was performed using the RPKM fold method, which extracts the input signal from the test signal.

(f) Rank analysis

Rank analysis of sequence usage was performed using the RankProd package⁹⁵, in which the significance of the difference in ranks of the same elements is assessed and reported.

(g) DNaseI digital footprints generation

The DNase-Seq alignments were converted to DNaseI footprints using the uniquely aligned paired-end reads to determine the DNaseI cut sites within the chromatin (Hughes unpublished,³³). The density of these positions was calculated using a 3-bp window and converted into a UCSC bigwig file for viewing. Low frequency intervals correspond to protein binding sites where the DNA is protected from DNaseI cutting, while high frequency represent DNA regions accessible to DNaseI.

(h) DNase digital footprints calling

DNase digital footprints were called using the standard parameters of pyDNase Python package which implements the Wellington algorithm.

(i) DNase average footprints

The original algorithm for calculating average footprints was developed from the original digital footprints idea (Prof Jim Hughes), in collaboration with a colleague, Ron Schwessinger. The pipeline adapting this initial algorithm was developed and implemented by myself, to determine and quantify sequence usage in a specific cell line.

(j) Normalisation of average footprints

The average footprints were normalised using the naked DNA control. We calculated the pattern of cutting any 6bp motifs in genomic DNA and normalise the average footprints.

III. Results

Chapter I. Open-chromatin assays: optimisation and setup

(a) Introduction

The ability to isolate regulatory regions in an unbiased manner is extremely valuable in gene regulation as it provides insights into one of the main aspects of transcriptional regulation: the interaction of TFs with DNA. One of the features of the regulatory regions that has been exploited for their identification and annotations is their location in open-chromatin.

Open-chromatin is sensitive to the activity of the DNaseI endonuclease, and thus the regulatory regions can be isolated using a classical method, the DNaseI assay (described in Materials & Methods, Section 1.05 (a)). The regulatory regions have thus been called hypersensitive (HS) regions. As we isolated and characterise the regulatory regions based on their hypersensitivity, we will refer to them as HS regions for the rest of this thesis.

Technological progress has seen the DNaseI assay adapted from the visualisation of single sites when coupled with Southern blotting²⁸, to analysis of genome-wide hypersensitivity (DNase-Seq)^{33,96} or even base-pair resolution of TF binding sites (digital footprints). Despite the informative output, DNase-Seq is not as widely used as current protocols can still be limited in terms of cell-type specific optimisations, minimum cell numbers or sequencing depth required.

This chapter presents the development and optimisation of a new DNase-Seq protocol to allow for reliable and reproducible isolation of HS regions. We describe the setup of a newly published assay (ATAC-Seq) for isolation of open-chromatin regions using low cell number ⁴⁴. We compare the two methods at a genome-wide level and assess their output in erythroid cells. Finally, we generate footprints with the resulting improved DNase-Seq data and ATAC-Seq in order to study the TF binding at high resolution.

1.02 Assessing hypersensitivity using the DNaseI assay

(a) DNaseI assay and Southern blotting

The original DNaseI assay involves incubating cell nuclei with the highly processive endonuclease DNaseI, extraction of DNA and mapping of the HS sites by Southern blotting.

The recommended starting number of cells for the DNase assay is 5×10^7 live cells. Live cell nuclei are isolated by mild lysis of cellular membrane, split in 8-10 aliquots each containing an equal number of nuclei. Successive aliquots of cells are incubated with an increasing concentration of DNaseI so that, over the same incubation period, each aliquot represents a specific and increasing degree of digestion. DNaseI will generate double cuts in regions of open chromatin which will increase in frequency with the concentration of DNaseI, and will result in a gradual digestion of the DNA typically referred to as a “fade” (Figure 8).

This pattern of DNA digestion, the fade, is expected and desired in DNaseI assay as the gradient from non-digested fragments to digested allows for the visualisation of the HS fragments by Southern blot. In addition, as described later in this section, this gradient is necessary as it increases the probability of selecting the sample with the optimal degree of digestion for library preparation which can be highly variable between samples.

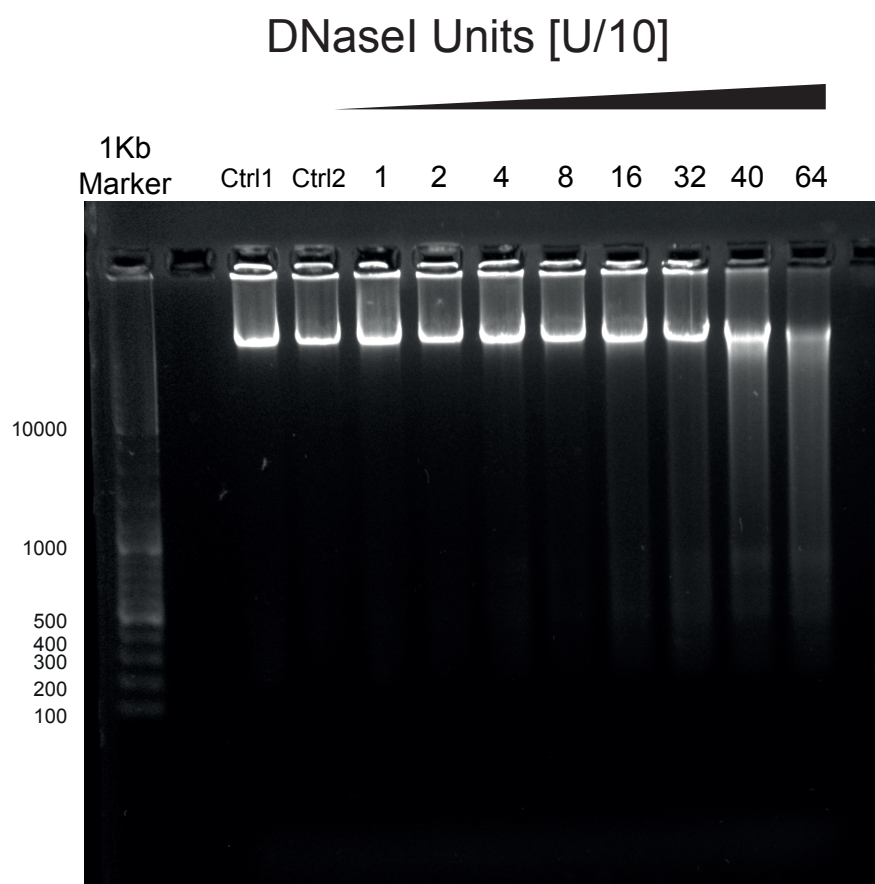


Figure 8. DNaseI digestion pattern ('fade'). Agarose gel electrophoresis (1%) of DNaseI digests of chromatin derived from human erythroid progenitors. Lanes 1 to 64 correspond to nuclei digested at 37°C for 3-5 minutes with increasing concentration of DNaseI (10, 20, 40, 80, 160, 320, 400, 640 U). Ctrl1 and Ctrl2 correspond to undigested samples (no enzyme) incubated on ice or at 37°C, respectively. Marker- 1Kb molecular weight marker (NEB).

(b) Known HSs can be visualised by Southern blotting

A specific DNaseI HS can be visualised using a Southern blot strategy as a decrease in the expected size of a restriction fragment containing the HS. The DNaseI digests are further digested with a restriction enzyme (RE), and Southern blotted. The hybridisation probe is designed to the ends of a RE fragment spanning the HS of interest.

In the regions where the HS was not cut by DNaseI, or it is below Southern blot detection, the entire RE spanning the HS is visible. If the DNA was sufficiently digested by DNaseI at the HS, smaller fragments representing the HS flanking sequences are detected.

An example of mapping the DNase I human regulatory element R1, corresponding to HS-48 can be seen in Figure 9 below. As expected, we see a band of 5.5 Kb in the control and samples with low amount of DNaseI. The 5.5Kb band intensity is reduced from DNaseI digest 16 to 64 indicating increased digestion. Two additional bands (2Kb and 3.5Kb) are visible in DNaseI digests ranging from 8 to 64.

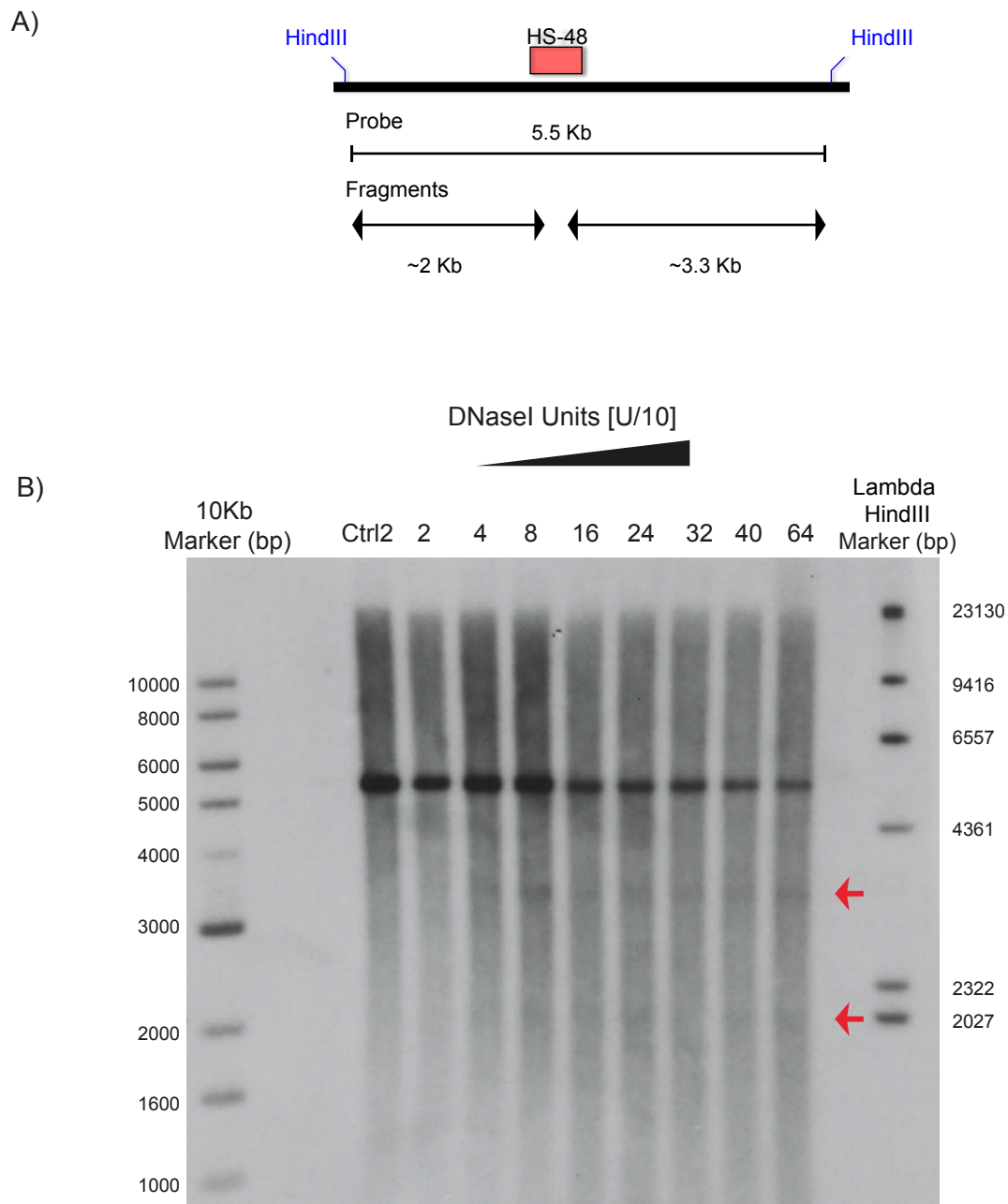


Figure 9. Mapping of human erythroid-specific DNaseI site R1 (HS-48). A) Representation of a 5.5 Kb fragment containing the R3/ HS-48 site, containing HindIII restriction sites, as well as the design of the hybridisation probe for Southern blotting. B) Example of autoradiograph used to identify HS-48 (Materials and Methods). Lanes correspond to samples digested with increasing concentration of DNaseI ranging from 0 to 640 [U] per sample and digested with HindIII. Following hybridisation, the entire HindIII fragment is visualised as a 5.5Kb band. For digests 8-64, HS-48 is detected and visualised in the form of two smaller fragments of ~2 and ~3.3 Kb (red arrows). Size markers used: 10 kb marker (left) and λ HindIII digest marker (right).

(c) Assessing hypersensitivity genome-wide using DNase-Seq

In order to analyse genome-wide hypersensitive sites, the digested DNA must be prepared for next generation sequencing using a library preparation protocol (Figure 10). The DNaseI assay generates 8-10 digests per sample analysed. In order to select which digests will be processed for sequencing, we decided to use the Southern blot results.

Thus, we hypothesised that the digests in which we can visualise a HS site at the model locus should have an sufficient degree of digestion to allow for genome-wide analysis.

Libraries were prepared from digests 40 and 64, as described in the Materials and Methods section. It has been shown that one essential part of the DNase-Seq protocol is enriching for small DNA fragments thought to contain HSs³². In order to select for smaller DNA fragments we reduced the length of PCR amplification step and performed a size selection for the 200-300bp range according to the manufacturer's instructions (NEBNext Library preparation).

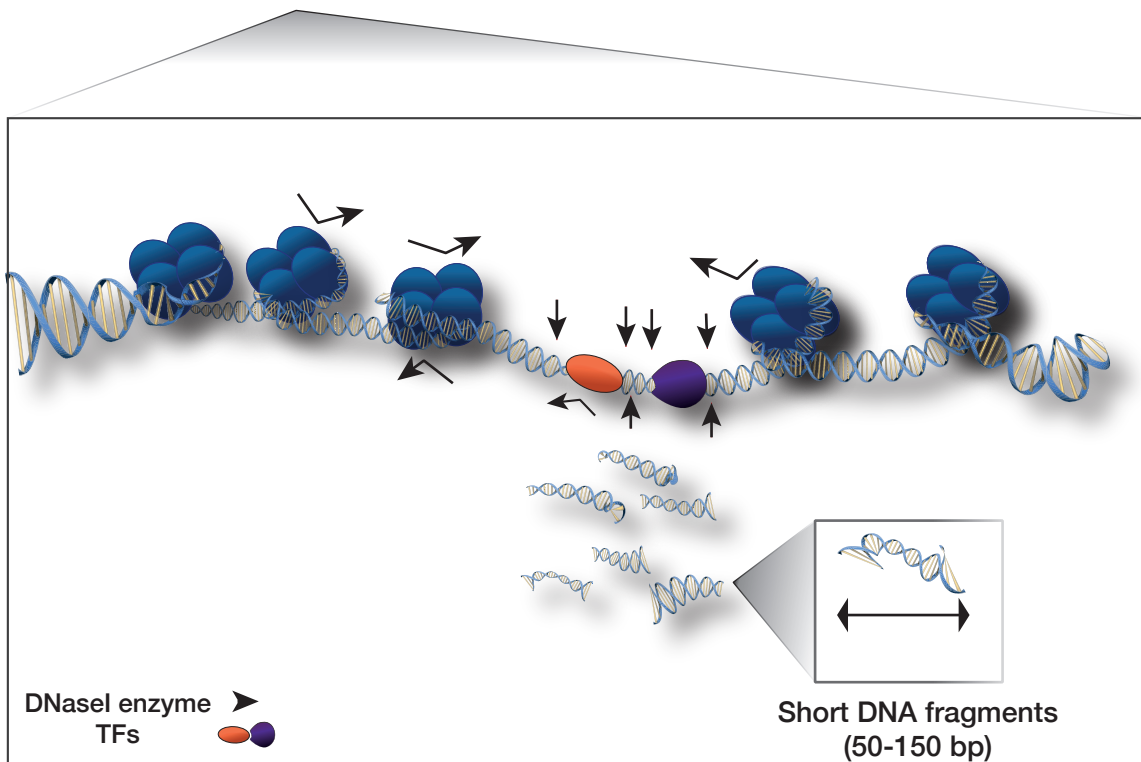
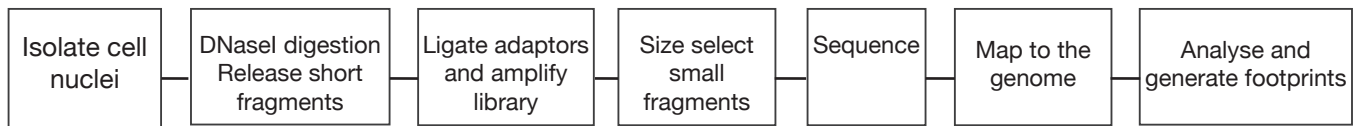


Figure 10. DNaseI assay and preparation of library for sequencing. DNaseI assay involves incubating live nuclei with the highly processive DNaseI endonuclease (black arrows) which will cut open chromatin preferentially. This process releases short hypersensitive fragments that can be processed for subsequent Southern blotting or sequencing.

Following sequencing, the samples were visualised in the UCSC browser along with the only publicly available erythroid DNase-Seq control from the K562 cell line.

The pooled library from digests 40 and 64 did not show any enrichment as compared to the K562 control (Figure 11). Background is visible in the test library, but no enrichment over expected peaks such as the erythroid specific genes (Hba) or the ubiquitous gene (Nprl3). The data set was analysed genome wide with no success in identifying enrichment at hypersensitive sites.

We set out to explain the discrepancy between the analysis of the DNaseI material by Southern Blot and by sequencing. We hypothesised that lack of visible enrichment (Figure 11) could be explained by two factors: 1) insufficient sequencing depth to distinguish signal from noise in these complex datasets, 2) high level of background within the sequencing assay.

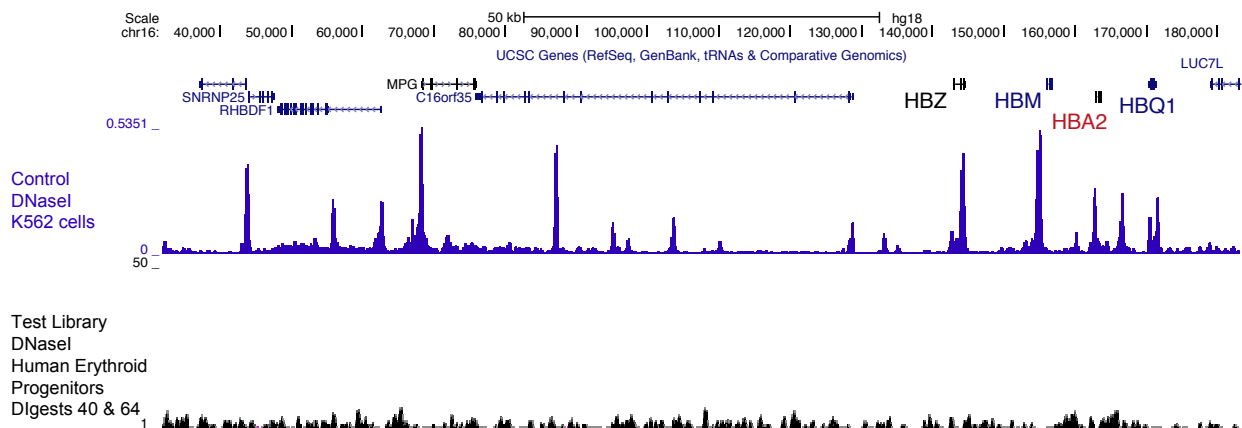


Figure 11. First DNase-Seq library for human erythroid progenitors. UCSC tracks showing the human α -globin cluster. In blue, the publicly available (ENCODE) human DNaseI from K562 cells (Control); in black the first library generated from digests 40 and 64 of the human erythroid progenitors DNaseI assay. Enrichment in open chromatin at HS regions as detected by the DNase-Seq can be seen in the form of peaks in the Control. In comparison, the test library does not show enrichment at the expected HSs. RefSeq genes are shown on top.

(d) Identifying sequencing depth sufficient for detection of HS signal

In order to test the effect of sequencing depth, we used an existing data set published from the lab ³⁴ and computationally subsampled this data to a low number of reads, comparable to the output of a MiSeq Illumina experiment. Sequencing on the MiSeq Illumina platform with the then current chemistry, would result in 18-20 million reads in total. Thus, the control dataset, sequenced on a GAII Illumina platform containing 40 million mapped reads, was randomly subsampled to 45% and 30% of the total reads (18 and 12 million reads respectively)(Figure 12).The level of enrichment is reduced (peak heights) in the subsampled data sets. However, the signal is distinct from the background. This computational experiment allowed us to conclude that even though the level of signal could be reduced, we should be able to detect HSs by sequencing on a MiSeq platform.

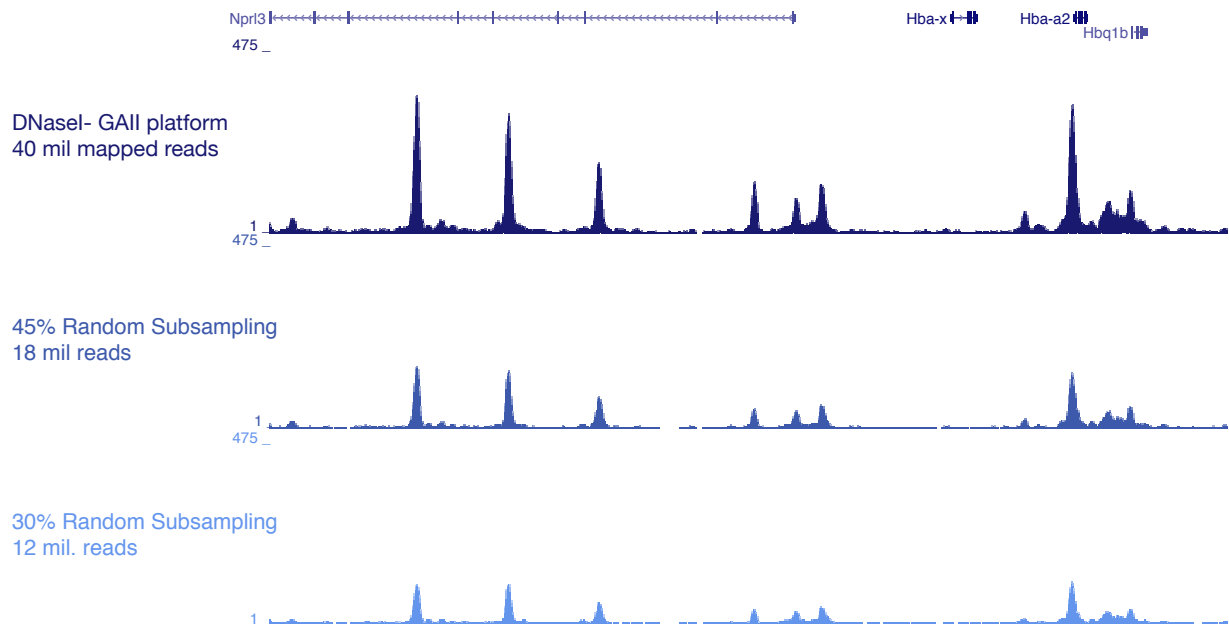


Figure 12. Random subsampling of DNaseI control sample. UCSC tracks displaying samples which were randomly subsampled from control sample to simulate the output of a MiSeq run in order to identify minimum reads required for visualisation of HSs. Subsampled data sets were generated using Samtools by random subsampling 45% and 30% of the control sample reads. The subsampled data sets show enrichment at expected HS sites at the α -globin cluster. RefSeq genes shown on top.

(e) Background levels affect the quality of enrichment in DNase-Seq

The lack of detectable enrichment can be attributed to low signal to noise ratio caused by high levels of background, poor isolation of the fragments of interest or a combination of the two.

The high levels of background observed in the sequencing experiment are likely to be explained by the level of background in the samples chosen for library preparation. Sample 40 and 64, even though showed strong signal for the hypersensitive site of interest, are likely to have had high background as well. This is due to the fact that the amounts of DNA resolved by Southern blotting⁸⁸ is considerably lower than what can be resolved in a sequencing experiment. Thus it is likely that the digests in which the tested HS is present also have high levels of unspecific digestion.

To achieve an optimal signal to noise in the sequencing assay, we considered two variables that are most likely to influence the enrichment: (1) the range of fragment sizes of digestion fragments selected for sequencing and (2) the optimal degree of digestion to size-select these fragments from. As there is a broad range of fragments that result from the DNaseI digestion, and several possible degrees of digestion to isolate the fragments from, the potential combination can be quite complex.

(i) Identification of DNase-Seq informative fragments size

A characteristic of DNaseI HS is the occurrence of numerous cleavage events over a short open-chromatin region, resulting in small DNA fragments. Taking

size into consideration was shown to improve sensitivity and specificity of the assay when coupled with tiling microarrays ³². In order to identify the fragment sizes that result in high signal in a sequencing experiment, we used an existing dataset (Ter119+ cells from mouse spleen) ³⁴, to identify the size ranges most enriched in HSs.

We extracted the mapped fragments and plotted their size distribution (Figure 13A). We then split the fragments into bins based on their fragment size, and mapped them back to the genome to be displayed in UCSC browser. This resulted in tracks containing only fragments of a certain size interval: 50-80bp, 80-110bp, 110-140bp, 140-170bp, 170-200bp, 200-230bp, 230-260bp (Figure13B).

The fragments in the library range from 50bp to 250bp, with the majority of fragments between 150-250bp. Despite the distribution maxima being around 200bp, when we analysed the level of signal in the different size bins we observed that the most informative fragments, accounting for the largest proportion of the signal, are the small fragments, 50-110bp. The medium-size range (110-170/200bp) also show enrichment, while 200-260bp fragments contain little or no signal.

In the light of this experiment, it became evident that the size selection we performed on the libraries based on the manufacturer's instructions for latest library preparation protocol (200-300bp DNA fragments) would omit the informative fragments. For future experiments, size selection was performed isolating the DNA fragments of 50-200 bp.

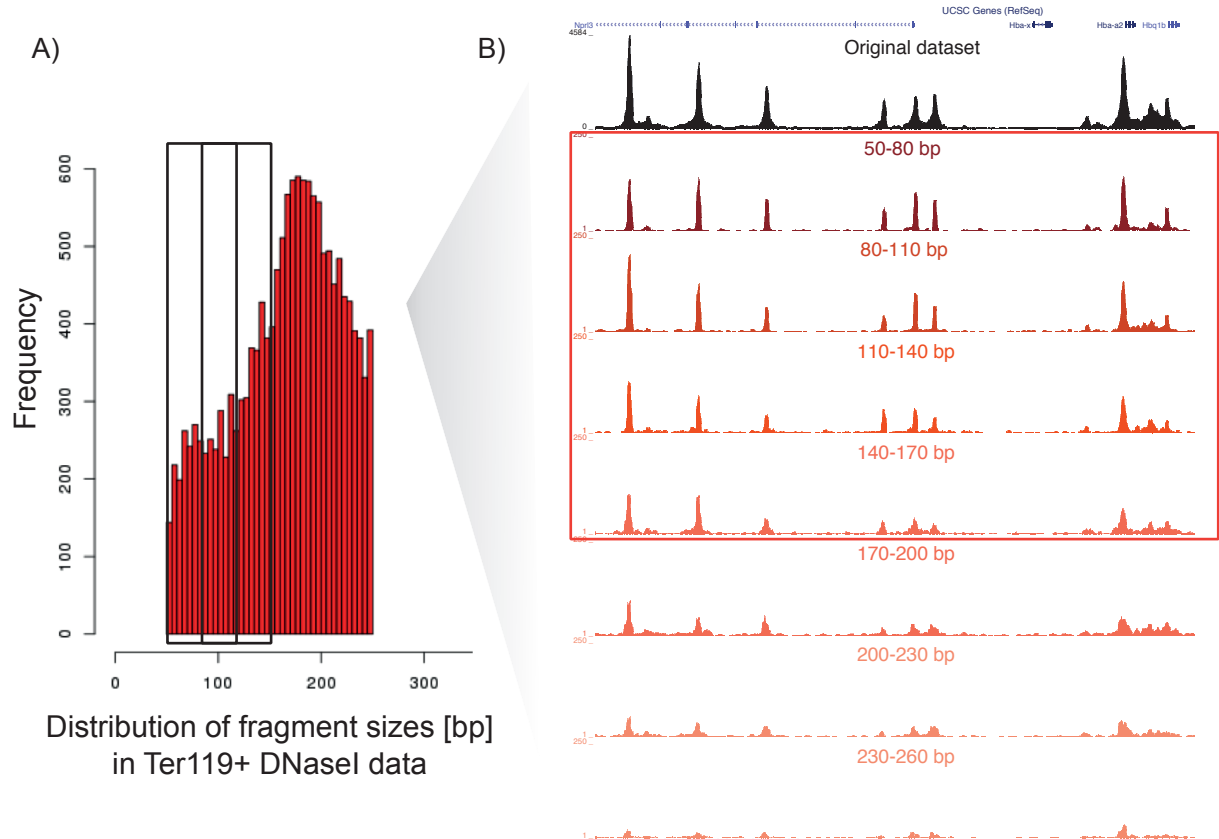


Figure 13. Identification of informative fragment sizes from existent data. A) Histogram showing the distribution of different fragment sizes in existent mouse (Ter119+ cells) DNase-Seq data. B) The mapped fragments were extracted from the data, split into different size bins and mapped back to the mouse genome. UCSC tracks of the different size bins, illustrating the fragments that contribute to the enrichment in the data (highlighted).

(ii) *Determining the optimal degree of DNaseI digestion for DNase-Seq*

The fragments we are interested in for sequencing, are not only small as determined in the section above, but are also present in very low amounts: the regulatory regions of the genome account for maximum 5-6% of the total DNA ⁹⁷. Thus, the degree of digestion of chromatin to release these fragments is essential for the quality of the signal. The DNaseI enzyme is an extremely processive enzyme that will, when present in large concentrations or for extensive time, cut DNA non-specifically, even in the context of nucleosomes ^{98,99}, generating in addition to the open chromatin regions also uninformative fragments that do not contain HSS- background fragments.

From our initial experiment (Figure 11) we have shown that Southern Blot analysis may not be a good reference for the optimum degree of digestion needed to produce libraries for sequencing probably due to over digestion (in addition to fragment size) . We set out to quantify the degree of digestion and identify the optimal parameters.

Considering the enzymatic performance of DNaseI can be a variable from assay to assay, a more quantitative assessment of digestion would help control for the variation between the samples of one experiment, and between experiments.

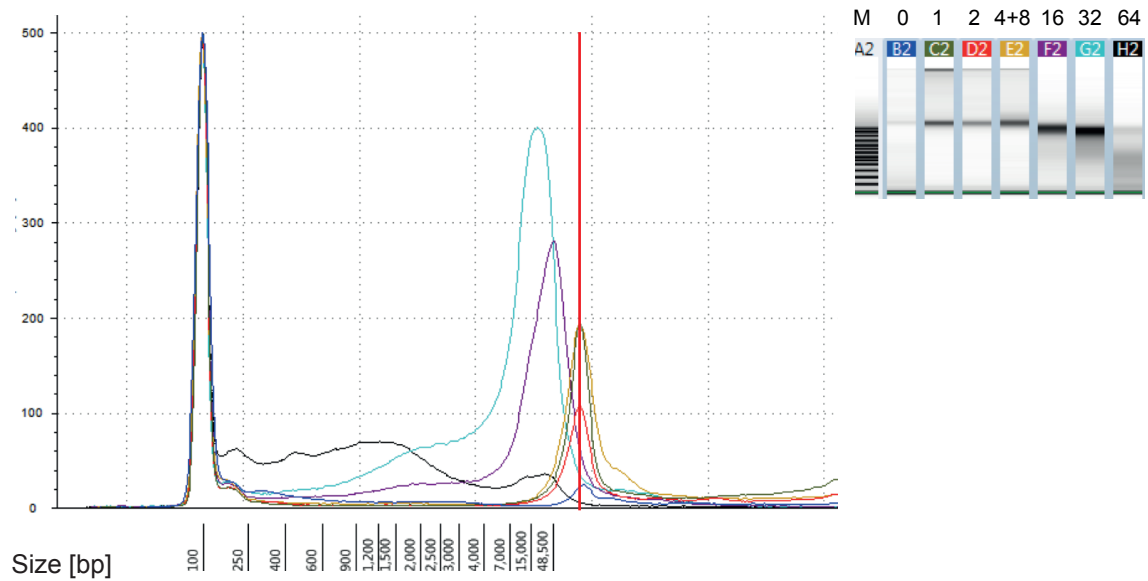
We explored visualising the low abundant fragments of interest on a platform more sensitive at detecting DNA than gel electrophoresis. One such platform is the “electrophoresis on chip” technology, performed using the TapeStation equipment (Agilent) which can reliably detect DNA amounts as low as 30ng.

It became apparent though, that even with increased sensitivity, it will be difficult to detect the changes in HS regions, the low abundant small fragments. However, the Tapestation can help determine another informative aspect which can be quantified: the ratio of very large genomic DNA fragments, to medium fragments and to relatively small. This ratio changes between samples and can be used to reconstruct the digestion process to a certain extent. (Figure 14)

All the digests from the human fade analysed by Southern blotting were analysed in this manner using the Tapestation. The fragment size histograms show that all samples contain a proportion of large fragments (>45000 bp), indicated by the peak at that size. The summits of the sample peaks 1-8 are aligned with that of the undigested control, indicating low degrees of digestion. Interestingly, sample 16 (light blue) and 32 (purple) show a shift in the peak corresponding to large fragments suggesting increased digestion. Sample 64 (black) shows a complete change of size distribution, with most of the fragments under 15000 bases, suggesting the DNaseI cutting has become unspecific and has digested a large proportion of the genomic DNA.

These ratios were quantified and we observed a steep increase of medium fragments between samples 4+8 and sample 16. We hypothesised this is indicative of a transition from specific to unspecific cutting of the DNaseI enzyme and decided to test this hypothesis by selecting digests 2, 4, 8, 16 for sequencing.

A) Tapestation profiles of individual digests



B) Quantification of fragments sizes in individual digests

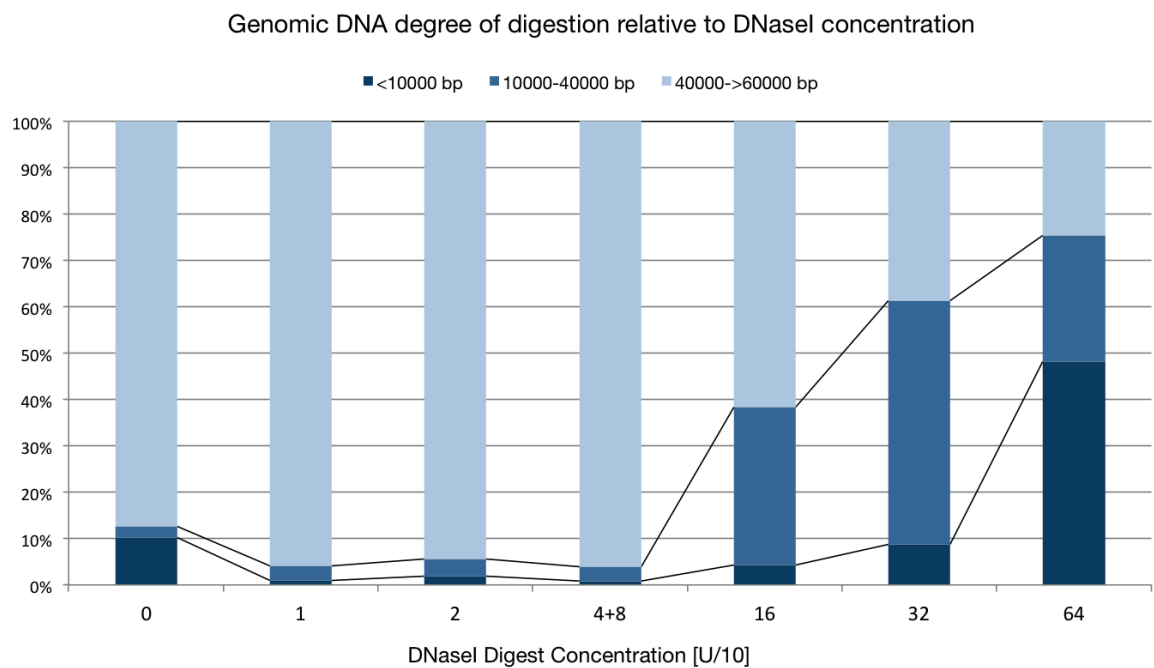


Figure 14. Assessing the degree of digestion in DNaseI digests. A) Tapestation profiles show histograms of DNA fragments in the different fragments analysed. The legend of the digests analysed can be seen on the left in a representation of the electrophoresis on chip similar to electrophoresis on gel. B) Fragment sizes were split into fractions: small (<10k), medium (10k-40k) and large fragments (>40k) and the percentage of each fraction was calculated for individual digests

(f) Correlating degree of digestion with level of enrichment in the data

The sequencing results of the four digests (2,4,8,16) (Figure 15) show that the best signal to noise ratio is detected in digest 4 and digest 8. Digest 2 has enrichment, but the level of background is higher than in digests 4 and 8. As we would predict digest 16 is overdigested, with high levels of background. This result shows that optimal DNaseI digestion is crucial for improving the enrichment in DNase-Seq assay. Tapestation profiles provide a good measure for the degree of DNaseI digestion and can be used as a quality control for determining which DNaseI digests can be processed for DNase-Seq.

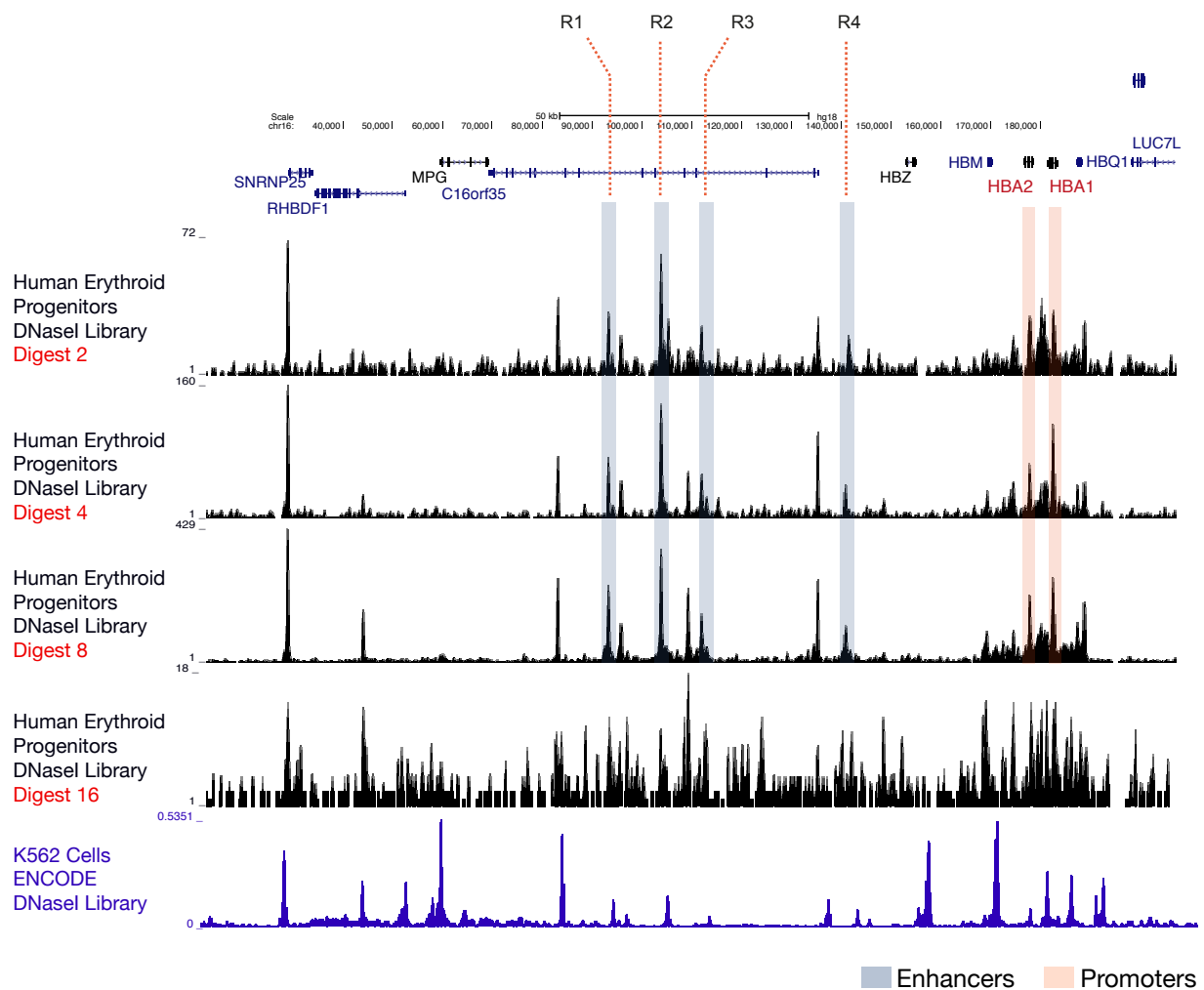


Figure 15. DNase digests show different levels of enrichment at HSs. UCSC tracks showing the the open-chromatin landscape of the human α -globin cluster in samples with different DNase concentrations. Libraries from human progenitor cells digests (black) and K562 control data set (blue) from ENCODE. RefSeq genes are show on top.

(g) Improved DNase-Seq protocol used to isolate HS regions in erythroid cells from a different species- mouse

Integrating selection of the most informative fragment size and of the optimal DNaseI digest in the DNase-Seq protocol resulted in high-quality, reliable genome-wide profile of HS regions in human primary erythroid cells. We next tested this optimised protocol on mouse primary erythroid cells, as the use of mouse models of erythropoiesis is essential for this study.

The DNaseI assay was performed on mouse Ter119+ cells purified from spleens of APH treated mice, a widely used source for adult erythroid cells (Materials and Methods). A DNase-Seq library (Test library) was prepared from the mouse erythroid cells using the optimised protocol, and sequenced on the MiSeq platform. This experiment resulted in the successful generating of genome-wide HS regions profile for primary mouse erythroid cells (Figure 16 A).

It is worth noting that this Test library has 2.5 million mapped reads as compared to 40 million mapped reads for the Control library (our reference sample for mouse primary erythroid cells ³⁴. Based on our subsampling experiment of this Control library (Figure 12 above) the signal level was expected to be low in samples with such low reads number. However, the Test library has much higher signal than predicted, suggesting that with the improved protocol we achieve better isolation of informative fragments. We confirmed this observation by analysing the size distribution of the fragments in the Test library (Figure 16B).

Thus, even though sequence at much lower depth than previously described, the quality of the signal is very high. This is very important from practical reasons, as it confirms that multiple samples can be multiplexed in one sequencing event, without compromising on the quality of the results. But most importantly, this experiment shows that our optimisations resulted in a more efficient DNase-Seq protocol that can be used to generate high-quality DNase-seq data, independent of the source of material and species.

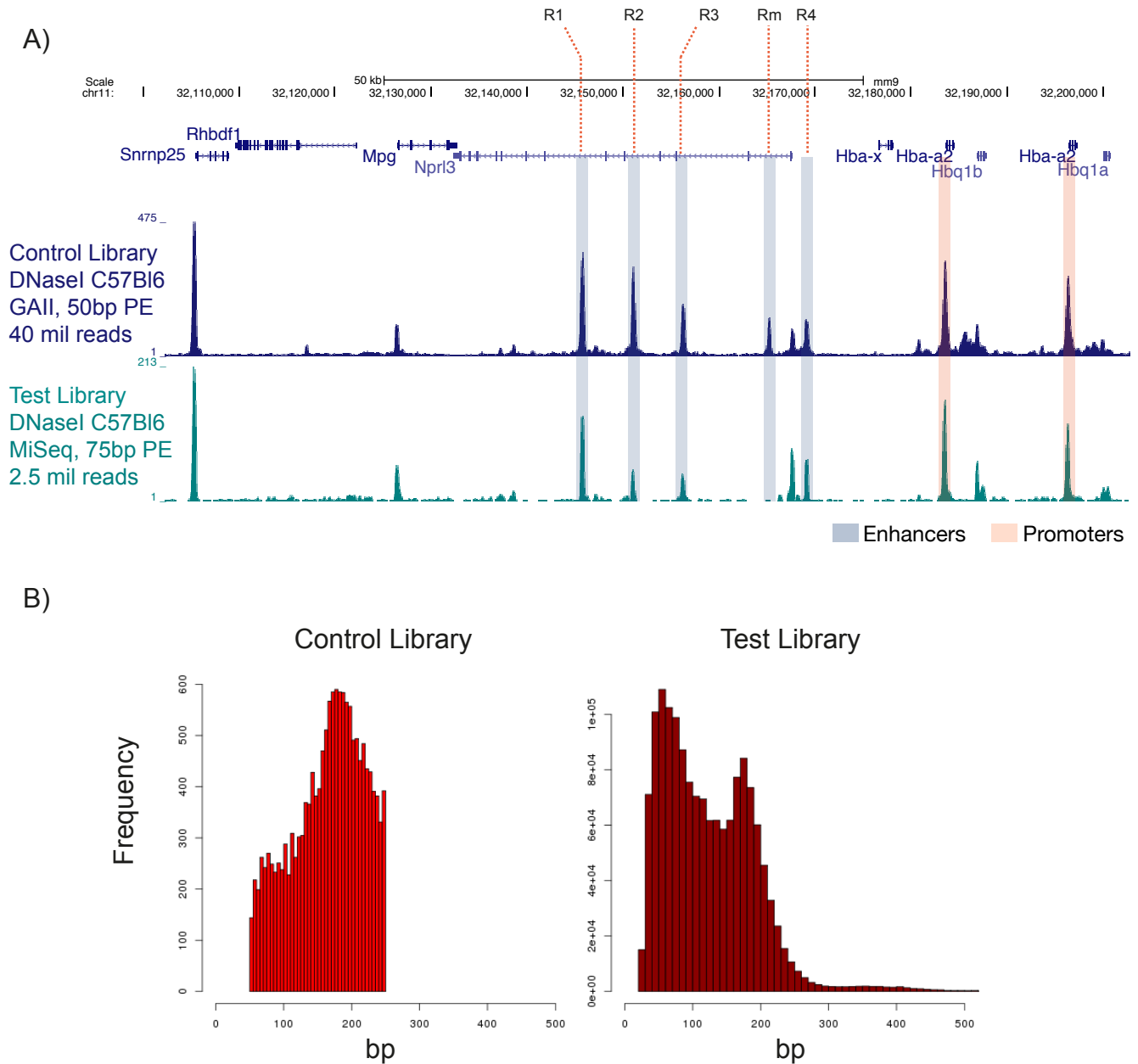


Figure 16. Isolation of HSs enrichment in mouse Ter119+ cells using the optimised protocol allows for identification of HS sites. A) UCSC tracks showing the the open-chromatin landscape of the mouse α -globin cluster in two different samples: Control library previously sequenced at high-depth on the Illumina platform GAI (dark blue), and test library generated with the optimised protocol, sequenced at low depth, on the MiSeq (light blue). The expected HS regions can be identified in the test library. B) Fragment size distribution in the libraries. The Test library contains a higher proportion of small fragments than the control library.

(h) Addressing the limitations of the DNase-Seq protocol

We have applied the improved DNase-Seq protocol in different cell types of mouse and human origin, as will be described in the next chapters. However, there are a few limitations that should be considered that make the use of this protocol prohibitive for early developmental and differentiation compartments, as well as for primary cells from patient samples.

These limitations refer to (1) cell type-specific optimisations; (2) duration of the protocol; and most importantly (3) cell number required.

Firstly, the cell type specific optimisations involve adaptations at the DNaseI assay stage because lysis and digestion of the DNA depend on nuclear compaction, size and level of nucleases in a cell. At the DNase-Seq stage as each cell type will have its own unique pattern of digestion that must be controlled for when preparing for sequencing.

Secondly, the preparation steps from nuclear isolation to a library ready for sequencing can take up to 3 weeks due to all the quality control steps involved. This is considered lengthy in a time when library preparations are becoming as rapid as one day procedures.

Thirdly, and probably most important, DNase-Seq requires extremely high cell input numbers. The recommended starting material is 5×10^7 cells. These amounts make studies of regulatory elements in primary cells from different developmental and differentiating stages extremely challenging.

A recent development in open-chromatin assaying addresses all these challenges. Assay for the transposition of active chromatin sequencing (ATAC-Seq)⁴⁴, is fast, does not appear to on cell type, but most importantly it can work on cell numbers as low as 500.

Considering our interest in studying early compartments in blood differentiation and development that are scarce in cell numbers, and based on our experience with improving the DNase-Seq protocol, we decide to setup and to test ATAC-Seq in erythroid cells.

(i) Open chromatin assay with low cell numbers- ATAC-Seq

ATAC-Seq is a method for isolating open chromatin regions using modified Tn5 transposase which can insert sequencing adapters into the native chromatin (Materials and Methods). The assay is a rapid two-step protocol and can use input material as low as 500-50000 cells⁴⁴. The initial part of the protocol is similar to DNase-Seq and consists of cell lysis and nuclei isolation. The nuclei are subsequently incubated at 37°C with the Tn5 transposases, which insert adaptors. For the library preparation step, DNA is purified and amplified. The resulting library can be checked on the TapeStation, quantified and sequenced.

The assay was performed as described in the Materials and Methods section, with input material of 60000-75000 Ter119+ cells/ replicate from APH treated spleen - same material used for DNase-Seq (Figure 17). The results show that enrichment is isolated reliably from the low input material and the open chromatin landscape at the locus of interest (alpha globin) is comparable to the pattern obtained from DNase-Seq. Thus, ATAC-Seq is reproducible in our hands as shown by the triplicates and good quality data can be obtained from cell numbers as low 50-60000/replicate.

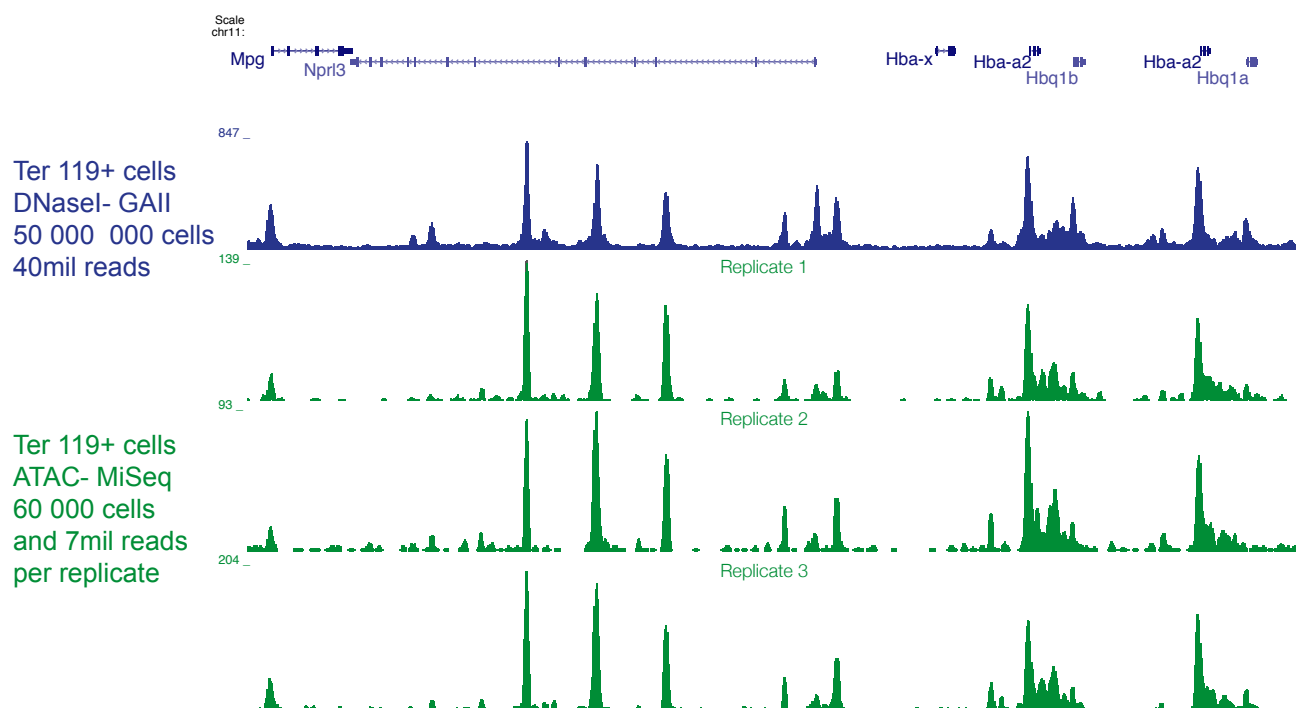


Figure 17. Isolation of HSs enrichment in mouse Ter119+ spleen cells using the low input method ATAC-Seq is extremely reproducible. UCSC tracks showing the open-chromatin landscape at the mouse α -globin cluster with two different methods: DNase-Seq, generated previously from 50 million cells sequenced at high-depth on the HiSeq (dark blue), ATAC-Seq, generated from 60000 cells/replicate sequenced at low depth, on the MiSeq (green).

(j) Genome-wide comparison of DNase-Seq and ATAC-Seq

As with any new method developed, the community is interested in the performance as compared to the current gold-standard(s). It is clear the ATAC-Seq offers practical advantages over the classical DNase-Seq. But do the assays perform comparably well? Are there any intrinsic biases that we should be aware of?

Having isolated the open-chromatin regions from the same cell type (APH treated spleens from C57bl6 mice), with both methods, we found ourselves in the unique position of being able to compare the two methods, not only at the locus of interest (described in the section above), but at a genome-wide level.

(k) Identification and characterisation of differential regions between DNase-Seq and ATAC-Seq

The data sets were peak called and the identified enriched regions were compared to isolate the peaks that are significantly different between DNase-Seq and ATAC-Seq.

(i) Differential analysis of DNase-Seq and ATAC-Seq enrichment

We must highlight that the differential analysis not a simple overlap analysis, but a robust quantitative assessment, in which we perform significance statistics.

An overlap analysis assigns the same relevance to small peaks, that are not statistically relevant but belong to one set, and to highly enriched peaks, which are statistically relevant but present in both sets. This because overlap analysis rely on peak callers decisions on occupancy in all regions, are not normalised, and thus are highly susceptible to factors such as sequencing depth which can reveal more peaks.

In contrast, a quantitative analysis, based on minimum three biological replicates, quantifies the enrichment (signal) at all identified peaks in say, DNase-Seq and ATAC-Seq, and assigns statistical significance. Thus, if a significantly enriched peak is unique to one set, it will be reflected by the statistics.

(ii) The ATAC-Seq peaks have significantly more enrichment than DNase-Seq

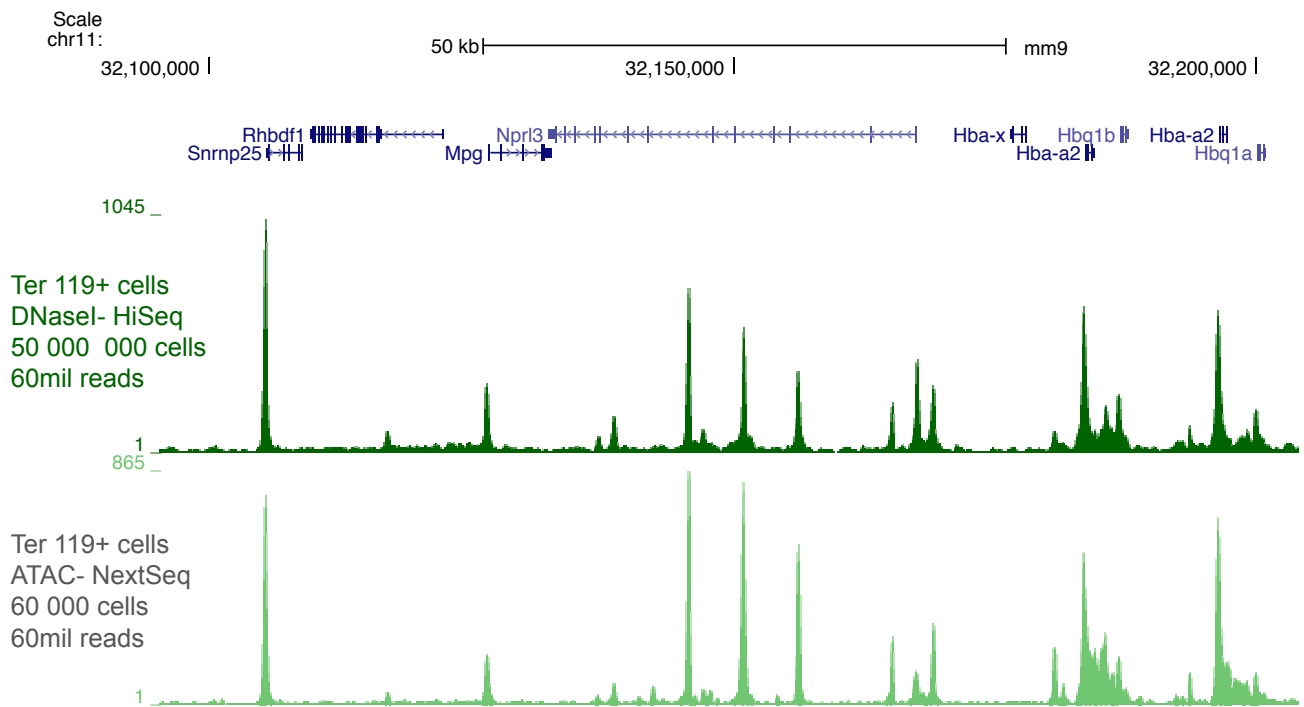
In order to identify the significantly enriched regions in each assay, we performed the differential analysis with DESeq2 (DiffBind package, Materials and Methods). It is an accepted practice to set a cutoffs for the level of fold difference that is reported, to distinguish the biological significant results in such analyses, from regions with low signal that tend to show higher variability¹⁰⁰. However, as we wanted to identify all the differences between the two assays, we have not applied any other cutoff except for the false discovery rate (FDR) of 0.01.

Out of the total peaks, 11968 (63.8%) do not show significant differences in the levels of enrichment between the two methods. This is comparable to the 77% overlap previously published with the ATAC-Seq method⁴⁴. Interestingly, the set

of significantly enriched peaks in DNase-Seq is rather low (4.2%), as compared to those enriched in ATAC-Seq (32%) (Figure 18).

In order to visualise the magnitude of the quantitative differences, the differential regions can be displayed as a mean-average plot (MAplot) which shows the level of enrichment of a peak relative to the fold difference between the two sets. Peaks enriched in DNase-Seq will have a positive log fold change, while peaks enriched in ATAC-Seq will have a negative fold change. Peaks that are present only in one of the data sets will have a higher fold change, whereas peaks that are enriched in one of the data sets will have a higher concentration, potentially combined with a higher fold (Figure 19).

A)



B)

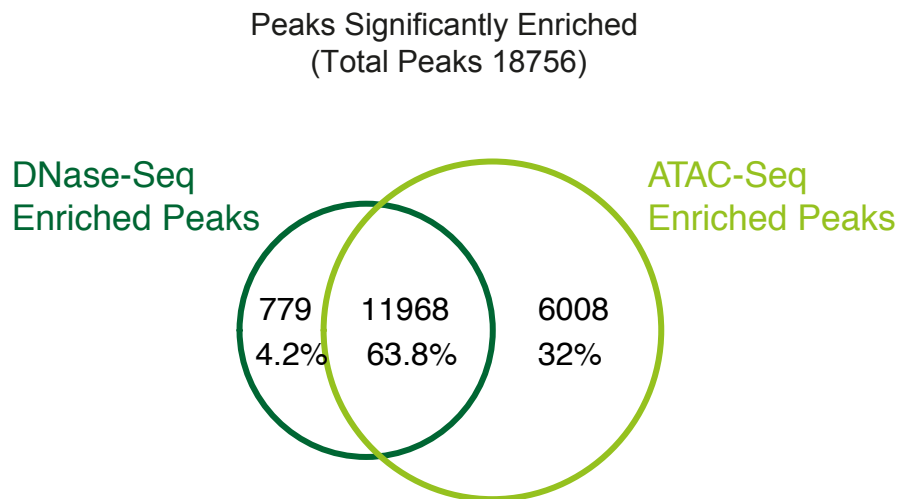


Figure 18. Comparison of DNase-Seq and ATAC-Seq in Ter119+ splenic erythroblasts. A) UCSC tracks showing the open-chromatin landscape at the mouse α -globin cluster with two different methods: DNase-Seq, generated previously from 50 million cells (dark green), ATAC-Seq, generated from 60000 cells/replicate (light green). The samples were normalised for sequencing depth. RefSeq genes are shown on top. B) DNase-Seq and ATAC-Seq peaks enrichment. Venn diagram illustrating the proportion of peaks, out of the total, that are differentially enriched (FDR < 0.01) in DNase-Seq and ATAC-Seq.

Differential Peaks: DNase-Seq (779) vs. ATAC-Seq (6008)
(6787 FDR < 0.010, from 3 biological replicates per assay)

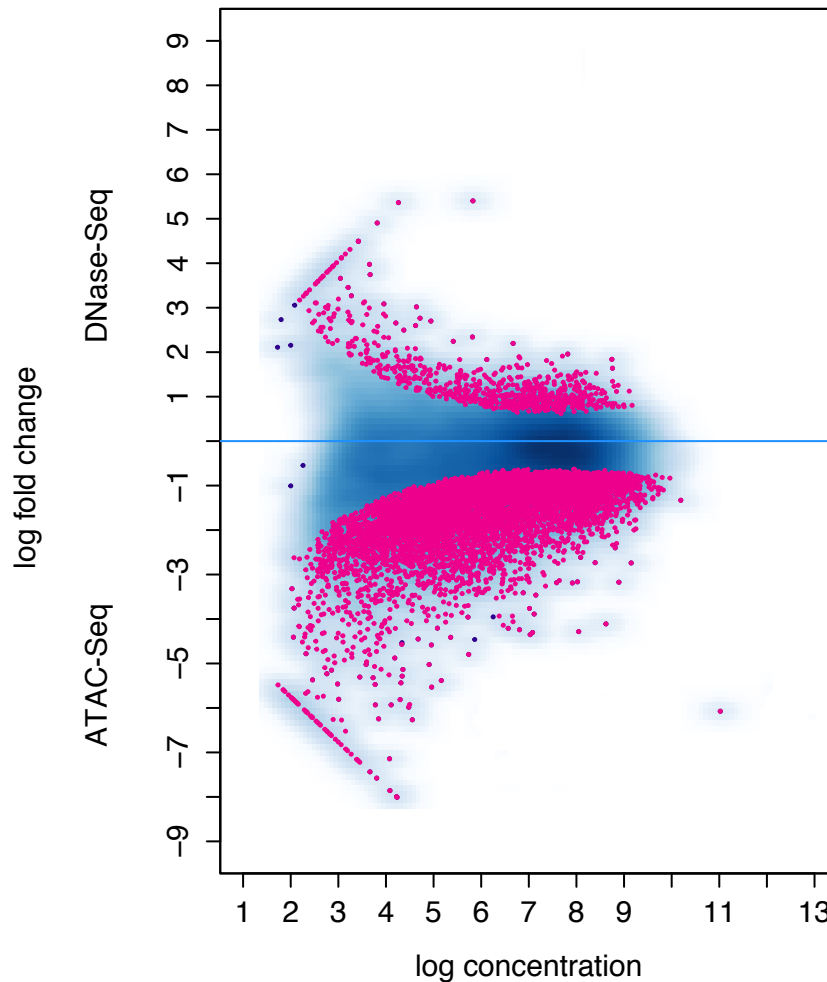


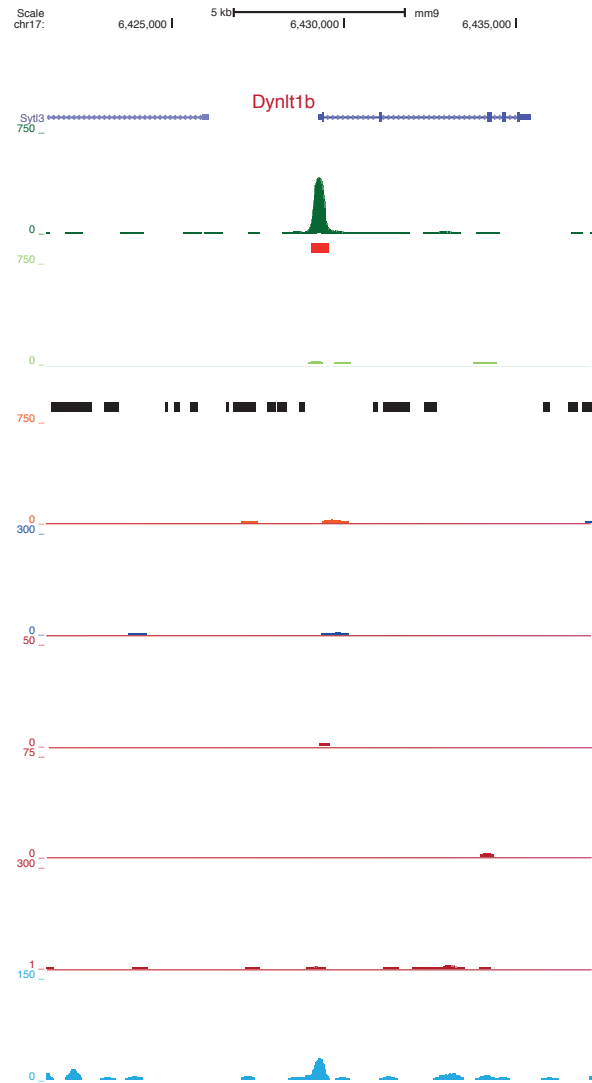
Figure 19: MAplot showing the differential peaks (FDR<0.01) between the DNase-Seq and ATAC-Seq. Fold changes in read density (log2) between two samples (y-axis) are plotted against the average of read density (x-axis) at the same peaks. Differential peaks in DNase-Seq will have a positive log fold change, while differential peaks in ATAC-Seq will have a negative fold change.

The majority of the quantitative changes in both DNase-Seq and ATAC-Seq show between 0 and 2 fold difference. Thus, even though there are more changes observed at the ATAC-Seq peaks, these are not significant and probably reflect higher signal at the same peaks. However, there is a set of regions with fold changes greater than 2. In order to relate fold changes with the signal in the data, we extracted a few examples with different significance thresholds.

(iii) Examples of regions significantly different between DNase-Seq and ATAC-Seq

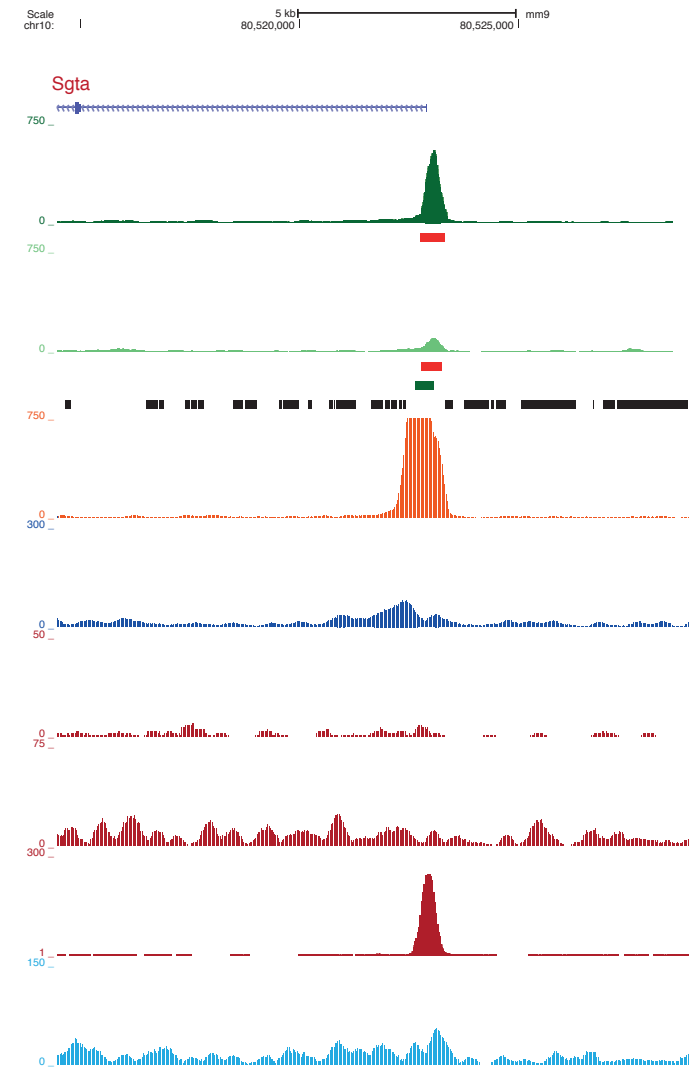
We could find only one region that can be identified only by DNase-Seq and not in ATAC-Seq, the promoter of the ubiquitous gene Dynein1b (Figure 20). The rest of the DNase-Seq differential regions are significantly enriched in DNase-Seq, but also present at reduced levels in ATAC-Seq. Thus, the differential DNase-Seq regions reflect variations in enrichment, and not the absence of ATAC-Seq peaks.

A)



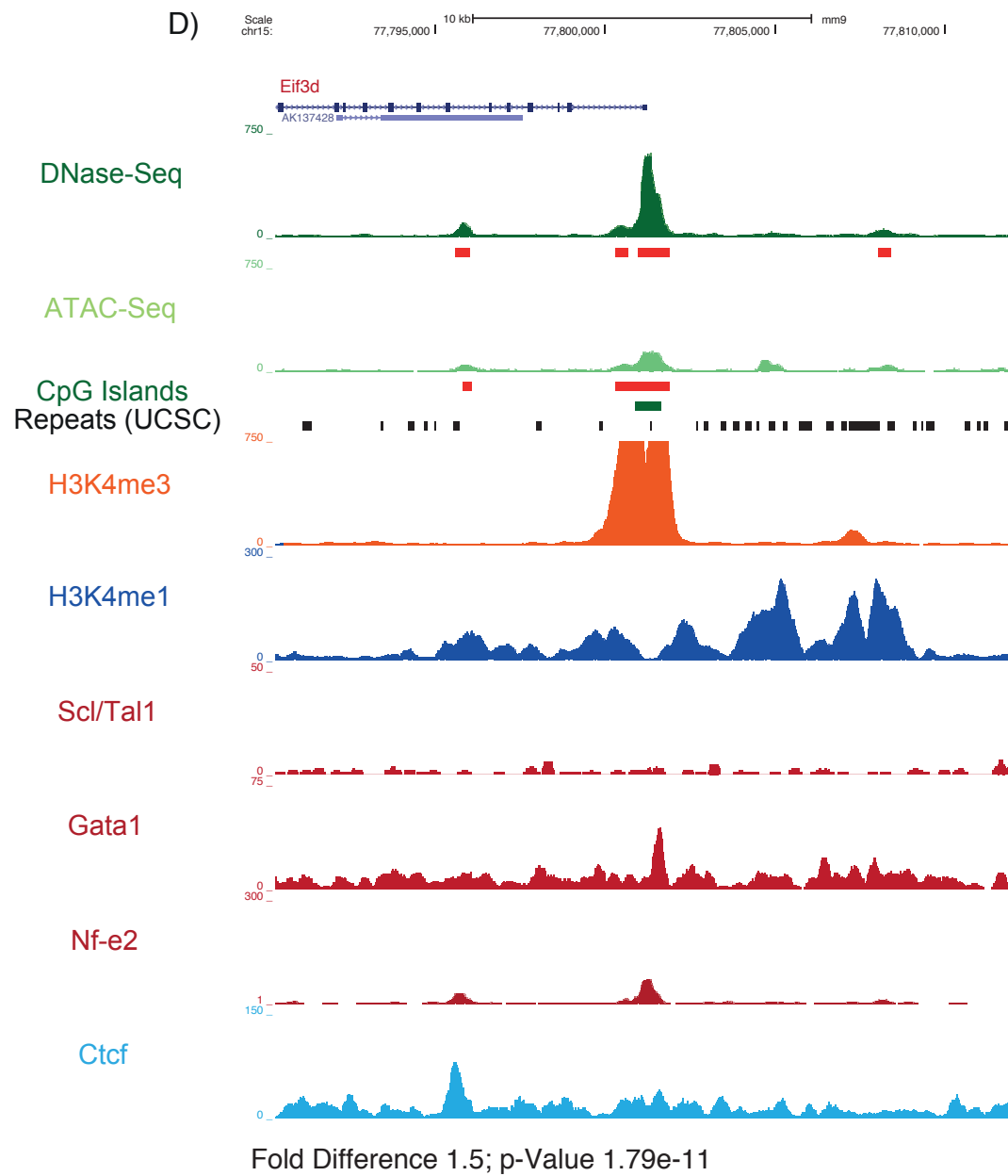
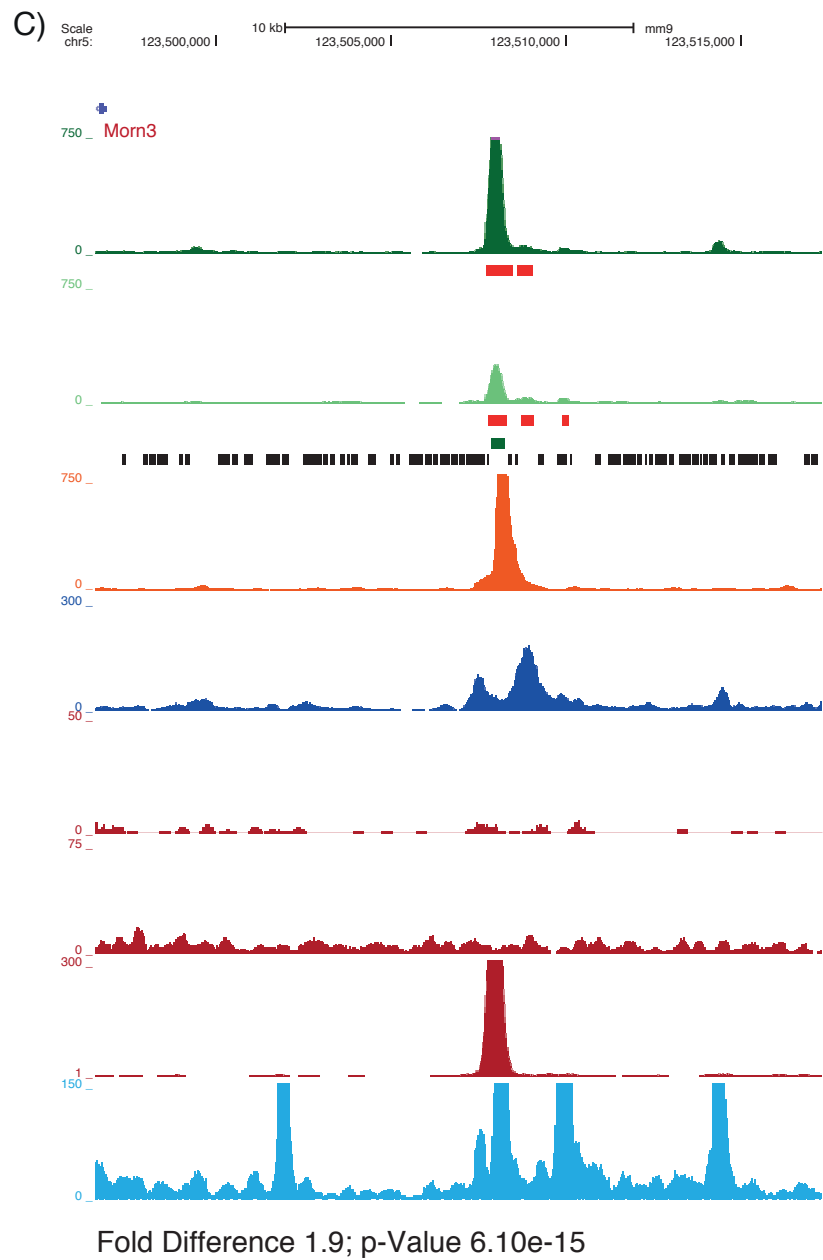
Fold Difference 5.4; p-Value 1.36e-26

B)



Fold Difference 2.2; p-Value 2.95e-16

Figure 20. Examples of statistically significant regions that are enriched in DNase-Seq, in definitive erythroid cells from APH spleen. UCSC tracks showing examples of four genomic loci where DNase-Seq signal is significantly higher than ATAC-Seq signal: A) *Dynlt1b*; B) *Sgta*; C) *Morn3*; D) *Elf3d*. Chromatin marks, erythroid TFs and CTCF tracks from ChIP-Seq generated previously in the lab are shown below the HS tracks. Ref Seq



The regions with the strongest fold differences in ATAC-Seq however, had extremely low signal or could not be detected in DNase-Seq. This is also reflected by the statistics as the regions from ATAC-Seq have lower p-values and higher fold differences (Figure 21). Thus, ATAC-Seq does reveal regions that are not identified by DNase-Seq.

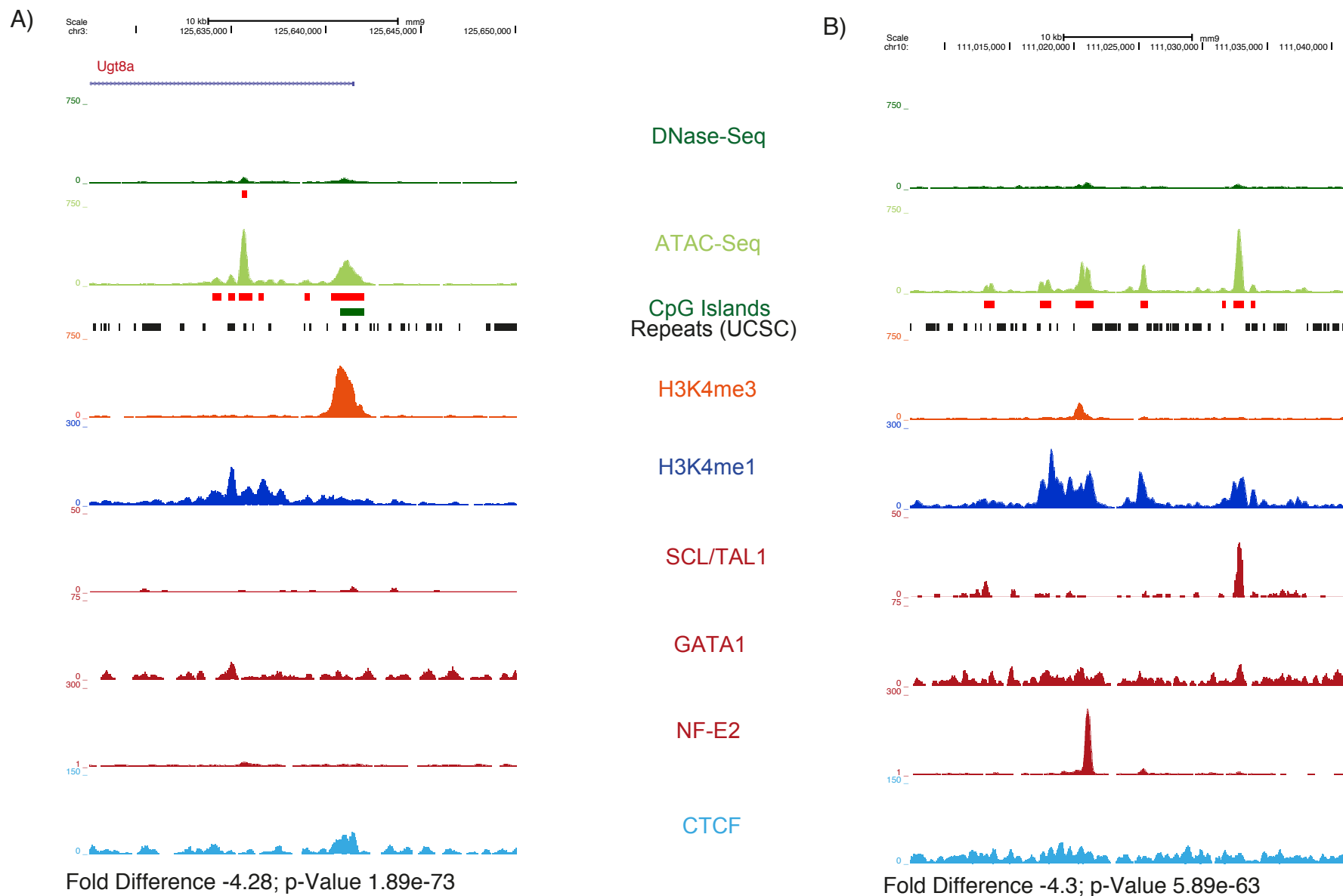
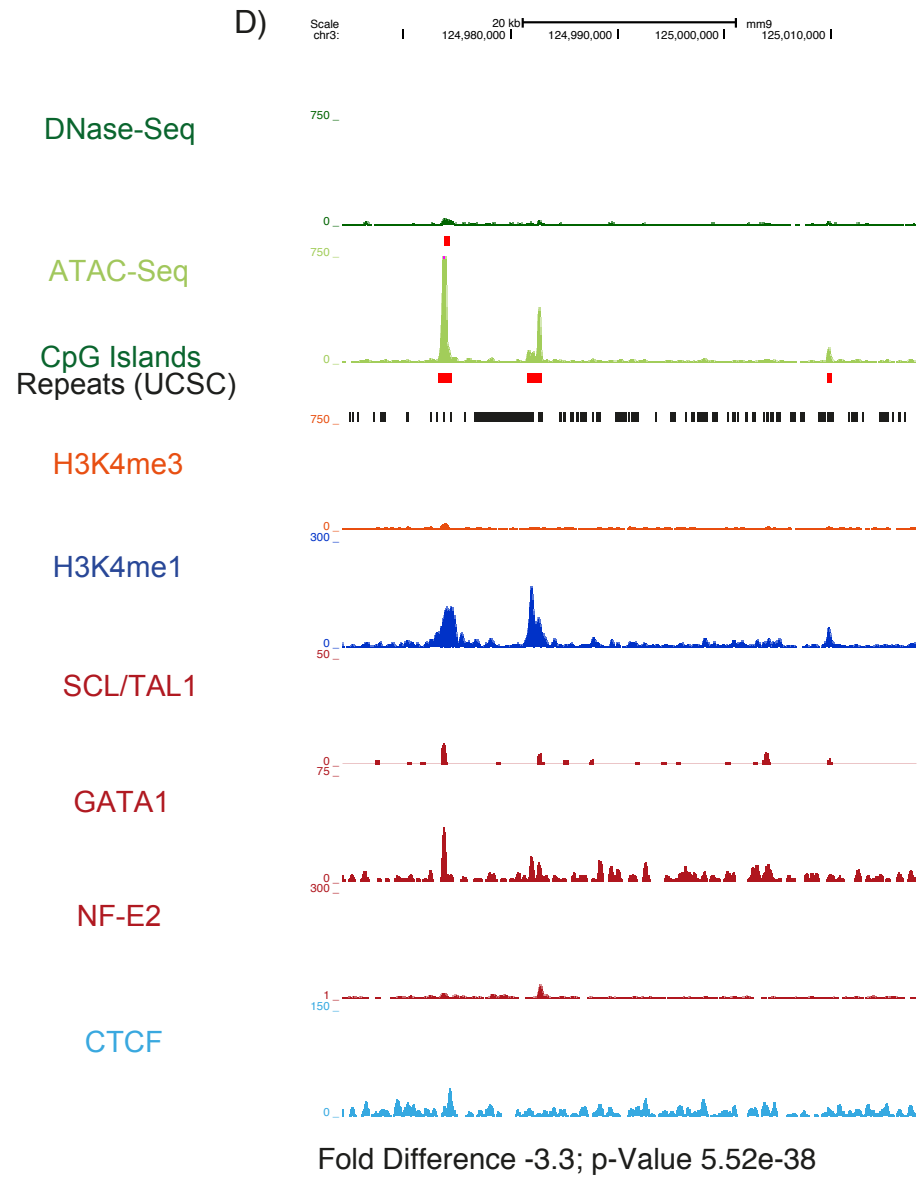
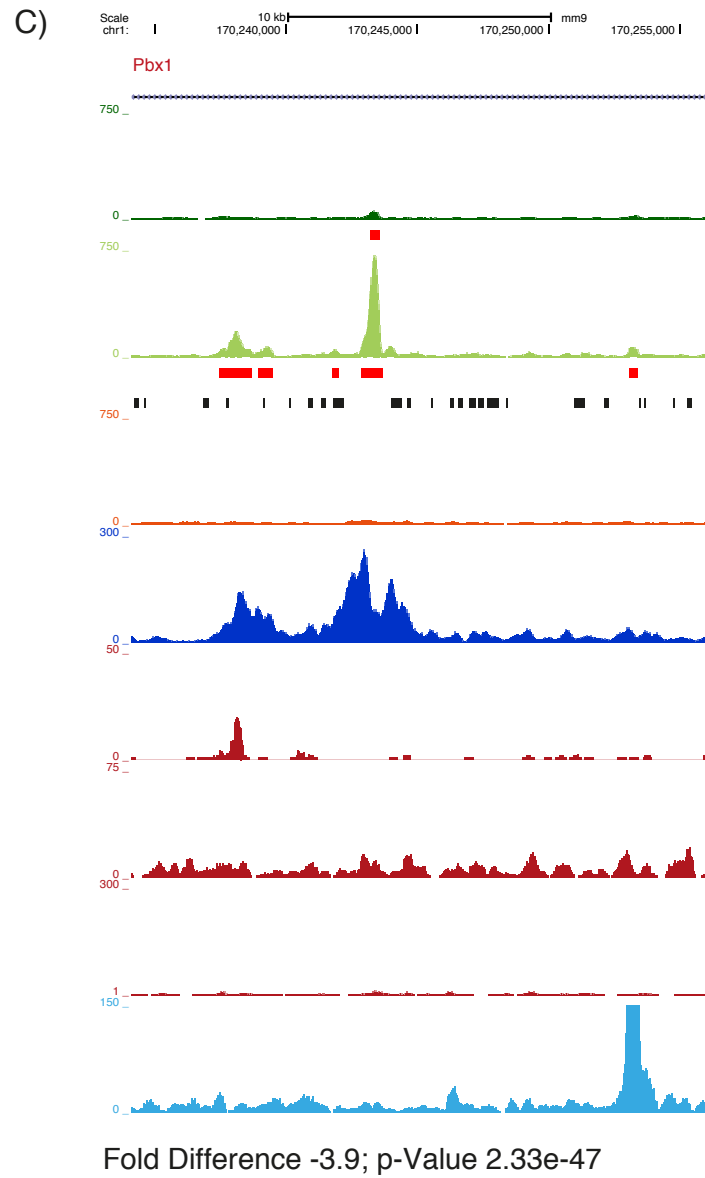


Figure 21. Examples of statistically significant regions that are enriched in ATAC-Seq, in definitive erythroid cells from APH spleen. UCSC tracks showing examples of four genomic loci where ATAC-Seq signal is significantly higher than ATAC-Seq signal: A) *Ugt8a*; B) Enhancer1; C) *Pbx1*; D) Enhancer2. Chromatin marks, erythroid TFs and CTCF tracks from ChIP-Seq generated previously in the lab are shown below the HS tracks. Ref Seq genes are shown on top.



(I) Annotation of differential regions

In order to identify which functional group the differential peaks from DNase-Seq and ATAC-Seq belong to, we annotated the significant peaks in terms of enrichment for chromatin marks and TFs.

(i) Enhancers and promoters distribution in the differential peaks

The annotation of the genomic elements was performed based on the data from Ter119+ cells, combining two criteria: difference of the chromatin marks H3K4me1 (associated with enhancers) and H3K4me3 (associated with promoters) enrichment¹⁰¹, combined with genomic location in terms of distance to TSS, intergenic, intronic or exonic - a combination of RefSeq and UCSC annotations. Elements that could not be annotated using any of these criteria were classified as “not annotated”.

The main observation from the annotation charts (Figure 22) is the difference in the ratio of promoters to enhancers in the differential peaks: DNase-Seq enriched regions consist of more promoters (56%) than enhancers (17%) as compared to ATAC-Seq regions, in which there is a more balanced distribution: 38% promoters and 33% enhancers. The proportion of elements not annotated is approximately the same: 22% and 23%.

Peaks Annotation

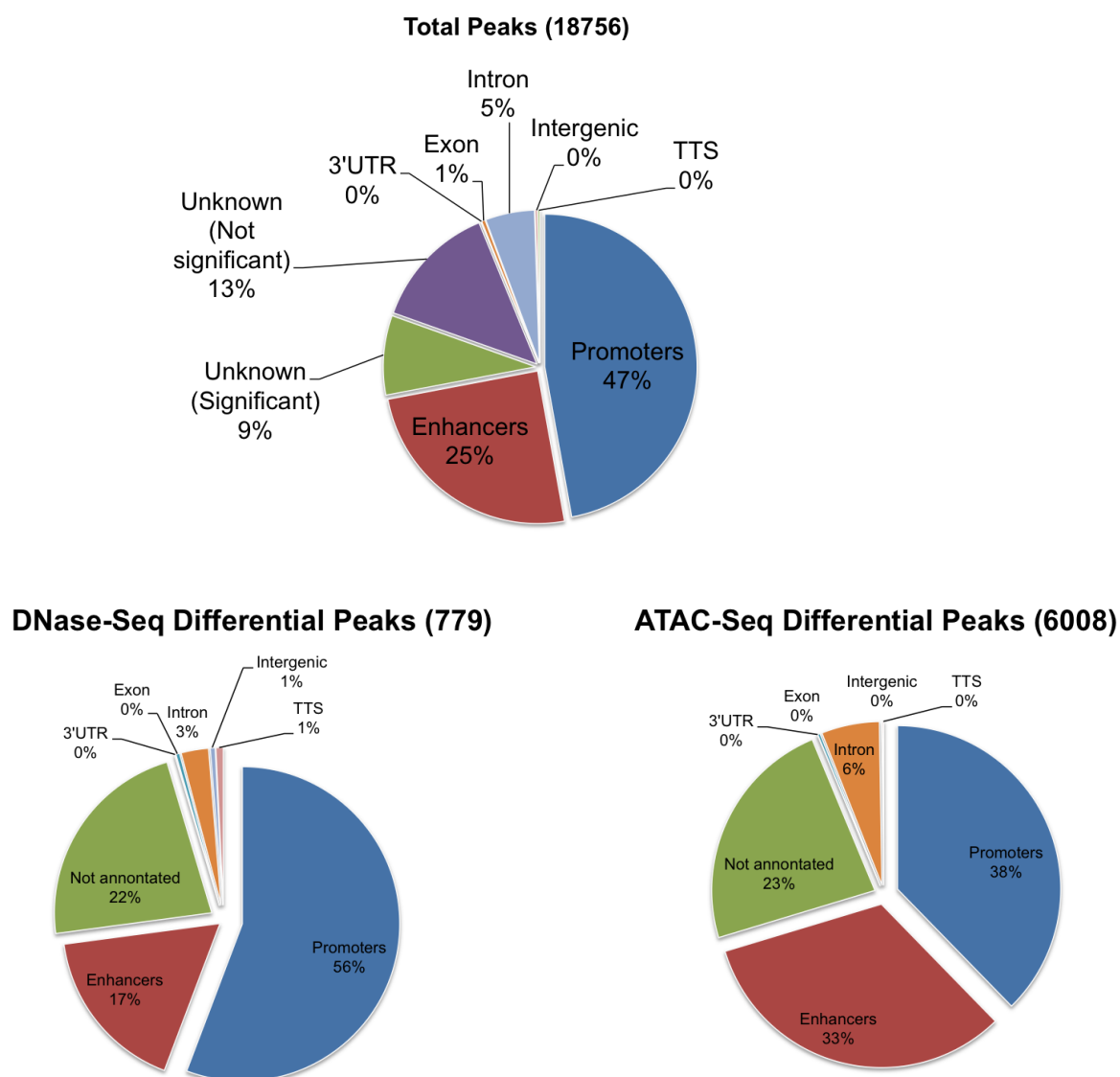


Figure 22. Peaks Annotation based on genomic location and chromatin marks H3K4me1/H3K4me3 ratio. Piecharts showing the annotation of total peaks analysed and of the differentially enriched peaks in DNase-Seq and ATAC-Seq. Enhancer and promoter elements were annotated based on H3K4me1/H3K4me3 ratio, as well as on genomic location from publicly available annotations. All the other elements were classified based on genomic location. The ‘Not annotated’ fraction could not be reliably identified based on chromatin marks ratio, or publicly available annotations (UCSC, RefSeq).

The distributions of promoters and enhancers as shown by the annotation charts, are recapitulated when plotting heatmaps of the H3K4me1/H3K4me3 ratio at the differential peaks (Figure 23). The regions identified as differential in DNase-Seq are consistently more enriched at promoters while peaks in ATAC-Seq are enriched for both enhancers and promoters. Another observation is that the signal at the ATAC-Seq peaks, in the total set, appears stronger than the signal of DNase-Seq, in agreement with the results from the differential analysis.

To summarise, there are a small proportion of DNase-Seq peaks significantly enriched, at promoters, and these are also detected in ATAC-Seq. ATAC-Seq on the other hand reveals unique peaks, with stronger enrichment at both enhancers and promoters.

As we are interested in using the assays for the study of transcription factor binding sites, we also analysed the binding of erythroid specific TFs and CTCF, using ChIP-Seq data generated previously in the lab.

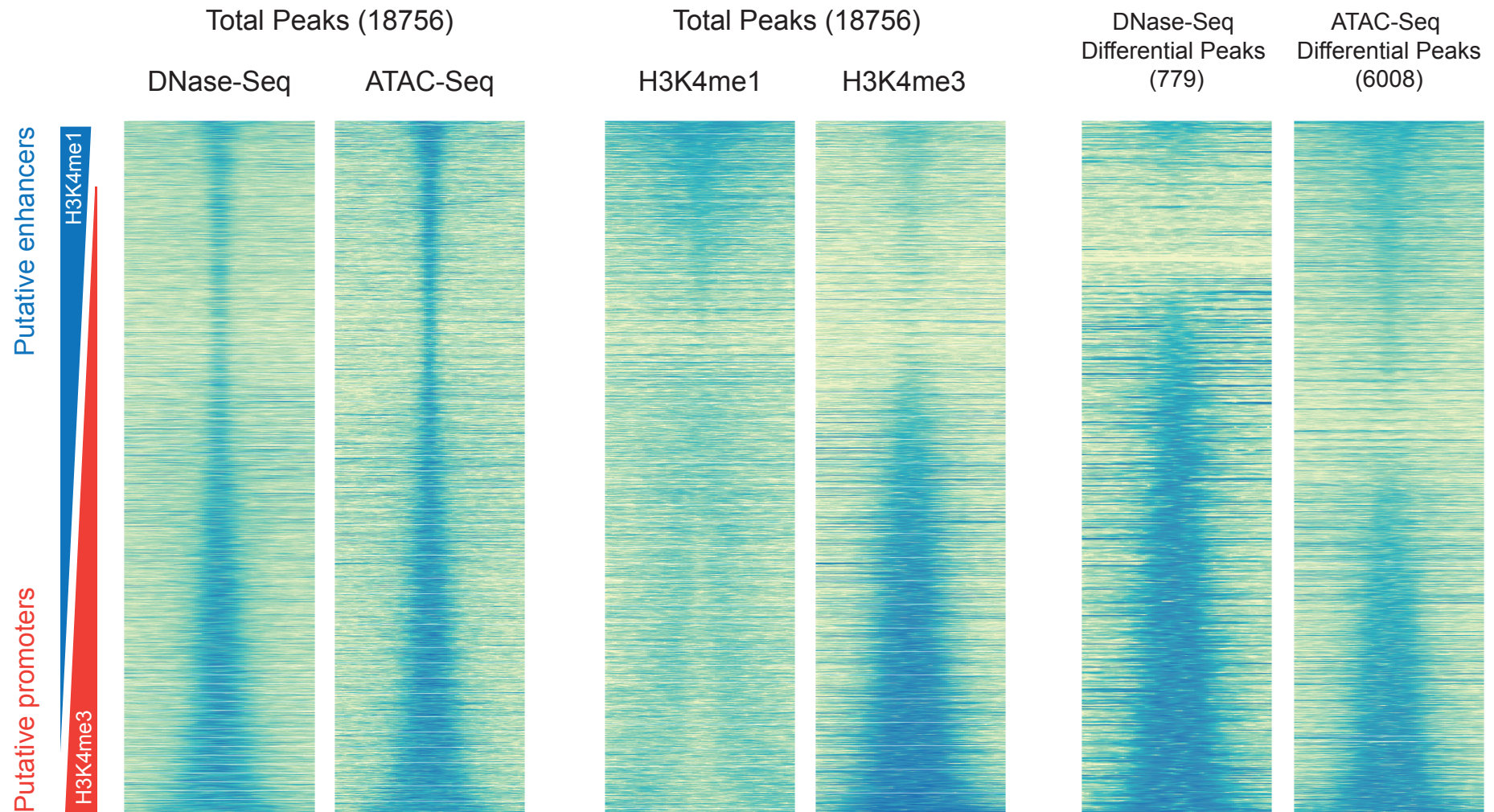


Figure 23. Heatmap representation of DNase-Seq, ATAC-Seq and ChIP-Seq signal 2kb around the total peaks or significantly differential peaks. Peaks are sorted by the difference in H3K4me1 and H3K4me3 enrichment. In the side panels the annotated "putative enhancers" are marked blue and "putative promoters" marked red. Enhancers and promoters are identified as hypersensitive regions with more than 10 reads per million (RPM) and a H3K4me1 to H3K4me3 difference in normalised reads higher than 1 ($H3K4me1 > H3K4me3$) or lower than 1 ($H3K4me3 < H3K4me1$) respectively.

(m) TF binding at the differential peaks

The factors analysed at the differential peaks are the erythroid TFs GATA1, NF-E2, SCL/TAL1 but also the ubiquitous boundary/insulator protein, CTCF. To analyse the level of enrichment at the differential peaks, as well as the total peaks, histograms of the coverage of DNase-Seq and ATAC-Seq and ChIP-Seq for the mentioned TFs from ChIP-Seq were generated (Figure 24).

The enrichment of ATAC-Seq and DNase-Seq at the total set of peaks confirms what we suspected from the differential analysis: the ATAC-Seq signal is higher than DNase-Seq at the same peaks (Figure 24A). In addition, as observed from the selected examples at DNase-Seq differential regions, the ATAC-Seq is more than 50% of the DNase-Seq signal, whereas at ATAC-Seq differential peaks the DNase-Seq is less than 30% of the enrichment.

In terms of TF binding, there is a visible increase of NF-E2 signal at the DNase-Seq differential regions, while ATAC-Seq differentials show low levels of SCL/TAL1 binding, and even lower of NF-E2 (Figure 24B). Another difference is the level of enrichment of CTCF binding in DNase-Seq enriched peaks. CTCF signal is almost at the same level with the DNase-Seq signal, suggesting that a large proportion of the DNase-Seq differential regions are CTCF elements. In contrast, ATAC-Seq differential regions show extremely low CTCF enrichment.

To summarise, DNase-Seq has generally lower signal at HSs, but at the differential regions is enriched for NF-E2 and CTCF; ATAC-Seq has consistently higher signal at HSs and at the differential regions it has low levels of enrichment for SCL/TAL1 and NF-E2.

A)

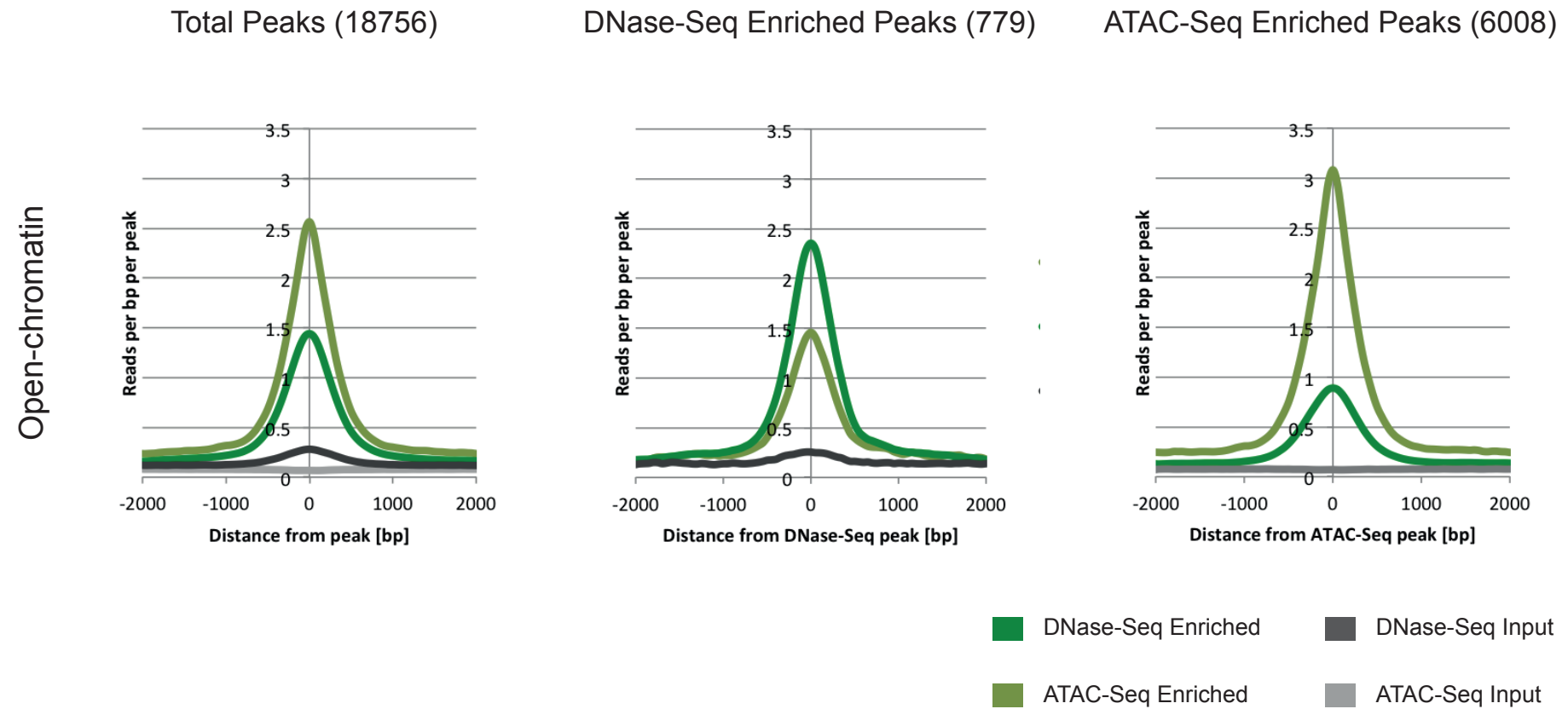
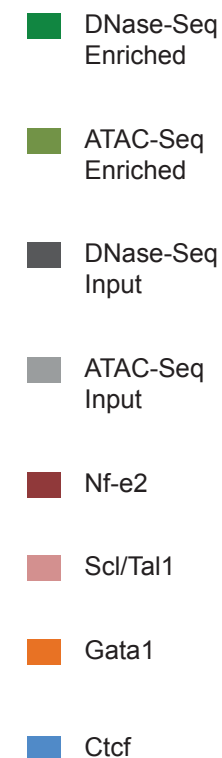
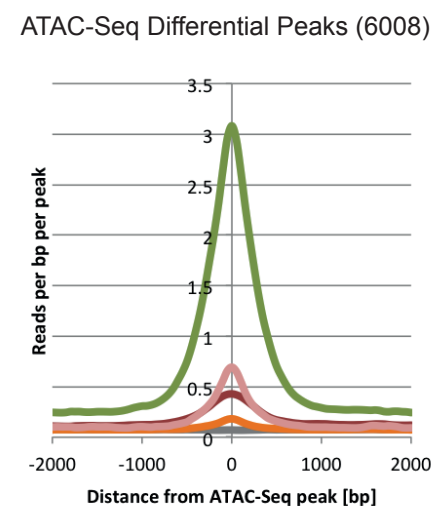
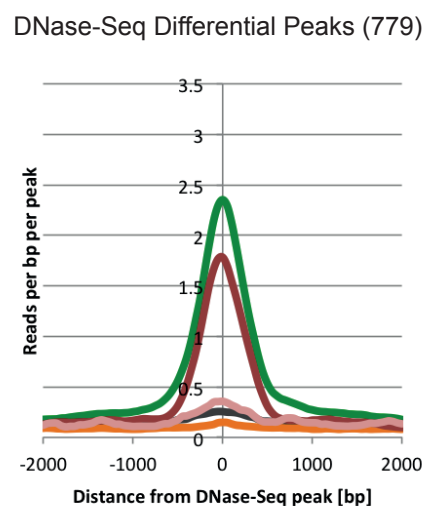
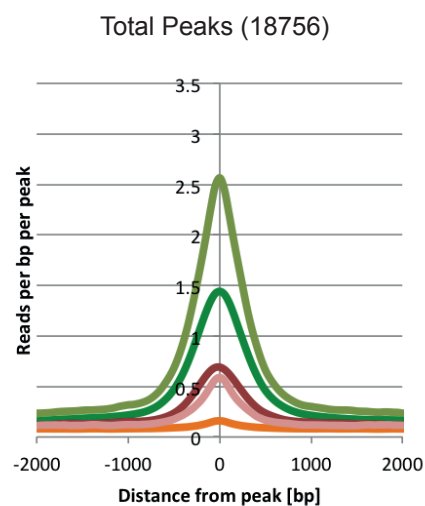


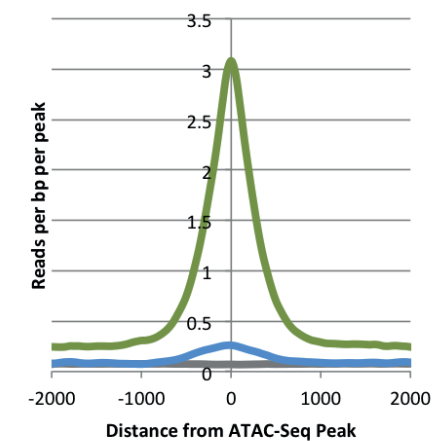
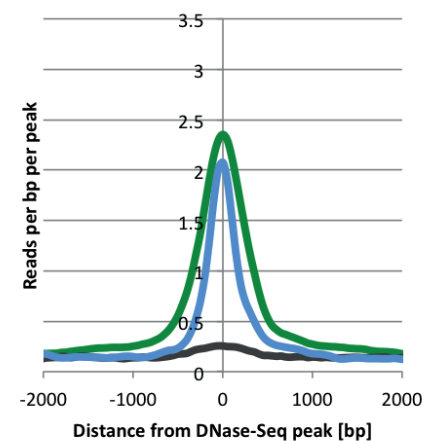
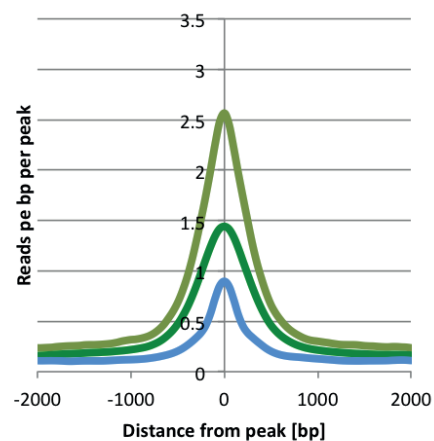
Figure 24. Histogram representation of signal across a 4Kb window centred on the hypersensitive sites analysed. A) Profiles of the open chromatin assays at the studied peaks, as well as signal from input material (naked DNA digested with DNaseI or Tn5). B) Profiles from ChIP-Seq of erythroid TFs as well as CTCF centred at the same peaks (total and differentials).

B)

Erythroid TFs



Ctcf



(n) Conclusion I

DNase-Seq is a powerful method for identification of open-chromatin regions, an aspect essential when studying gene regulations as these regions correspond to *cis*-acting elements. Understanding the properties of the sequences underlying *cis*-acting elements (short sequence, located in open-chromatin), helped us develop controls to standardise the method and improve the enrichment of open-chromatin.

ATAC-Seq is a new method that employs the modified transposase Tn5 to cut open chromatin and paste adaptors. ATAC-Seq offers practical advantages such as working on low cell numbers, and requiring little optimisation. However, in addition to the practical differences, we wanted to make sure we understand how the method performs, not only at a single locus, but also genome-wide. We thus compared DNase-Seq and ATAC-Seq from terminally differentiated erythroblasts from APH spleen (Ter119+ cells).

Overall it appears that the signal of ATAC-Seq is higher than that of DNase-Seq at the total peaks analysed. In terms of events we might miss if we only used ATAC-Seq, we would virtually miss no regions as the DNase-Seq differential peaks are detectable in ATAC-Seq. However, if we were to use only DNase-Seq, we would risk missing a big proportion of regions present in ATAC-Seq. These peaks are not only absent from DNase-Seq, but are also bound by erythroid TFs (NF-E2) suggesting they are probably enhancers.

As our main interest is to analyse TF binding at tissue specific enhancers, and these are not omitted by ATAC-Seq, but enriched, we can confidently use the new assay for assessing open-chromatin regions.

Before moving on to the subsequent chapters, there is one aspect that has not been discussed, and that is the high-resolution analysis (footprinting) of *cis*-regulatory elements. The ultimate aim of this thesis is the analysis of footprints formed by TF binding, thus footprint generation is essential. How do DNase footprints and ATAC footprints compare will be addressed in the last section of this chapter.

In the second part of the chapter we discuss high-resolution analysis of HS regions, digital footprinting.

1.03 Digital footprinting with DNase-Seq and ATAC-Seq data

One of the key steps leading to gene expression modulation is binding of TFs to regulatory regions. Identifying the exact TFBS within HS sites is critical for understanding aspects of gene regulation such as dynamics of factor occupancy or cooperativity, or how regulatory sequence variation can disrupt binding and lead to susceptibility to disease.

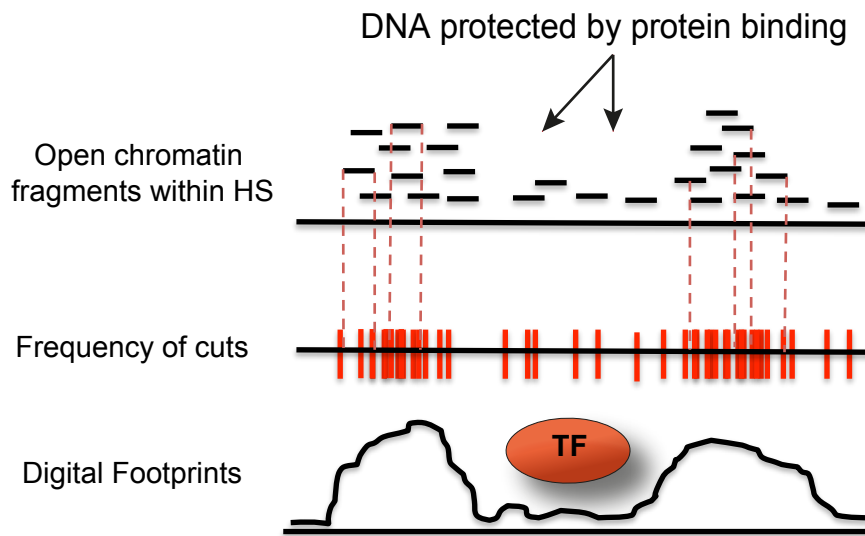
For the last three decades, DNase footprinting has been the method of choice for identifying TF binding sites through its various adaptations: as in vitro or in vivo chemical footprinting at a single locus ⁴², or as digital genomic footprinting at a genome-wide level ^{33,35,38}.

Generation of the digital footprints refers to the conversion of the low-resolution hypersensitive peaks into high resolution cut frequencies (Figure 25A). The analysis of the output of digital footprinting can be divided into two main steps: (1) display of the digital footprints and (2) computational detection (or “calling”) at a genome-wide level of the digital footprints.

This workflow is very similar to the analysis of ChIP-Seq data: the peaks are first displayed and then detected at a genome-wide level using peak callers.

In this section we will discuss the methods developed and tested for displaying and calling digital footprints. Secondly, we generate and compare DNase-Seq digital footprints and ATAC-Seq digital footprints.

A)



B)

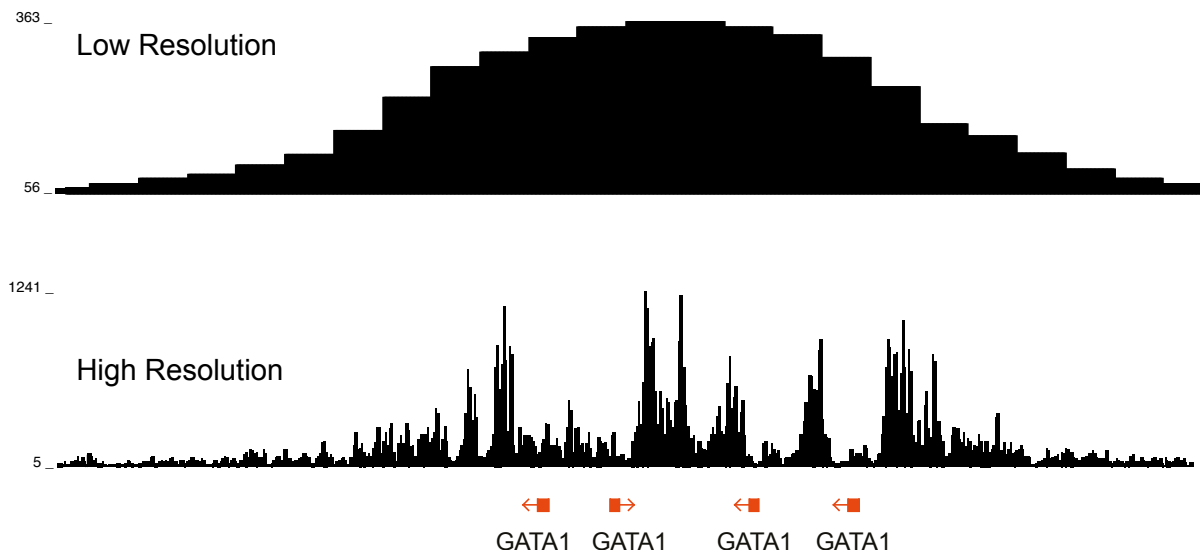


Figure 25. Illustration of how DNase-Seq data can be analysed to reveal digital footprints. A) Schematic representation of DNase-Seq cut locations as inferred from fragment ends: the cutting has high frequency in areas that are not bound by proteins, and low frequency in the area that are protected by proteins bound. B) HS region from erythroid data where the digital footprinting algorithm was applied, revealing at high resolution GATA1 binding sites protected from DNaseI cutting.

(a) Digital footprinting analysis

(i) Display of digital footprints

We believe displaying such high-resolution data should be a part of digital footprinting analysis. This is not only for the benefit of the user who can explore the data and formulate hypothesis but also for quality-control purposes. However, the challenges of displaying footprints are the pre-processing and the computational requirement to generate the high-resolution profiles. The already published approaches do not include the software for generating digital footprints tracks.

We developed an algorithm in house that allows for displaying of high-resolution footprints, and can be browsed in UCSC.

For the initial analysis in the current thesis, we use our in-house algorithm to visualise the footprint track. While visual assessment appears sufficient in some instances, for robust results we also used the pyDNase algorithm (described below) to computationally detect the footprints observed.

(ii) Detection of digital footprints

The accurate generation of digital footprints relies on correct modelling of the DNaseI mechanism. The endonuclease has a preference for open-chromatin, resulting in lower activity in regions protected by histones or protein-DNA interactions. In order to detect the digital footprints, the algorithms must search

for a depletion of DNaseI cuts which is surrounded by regions with high number of cuts (termed “ footprint shoulders”).

Approaches to detect footprints have pioneered in yeast³⁸(followed by development in mammalian cells^{33,96,102,103}.

The method developed by Boyle et al., 2011, to identify “digital genomic footprints”, DGF, is an improvement of the yeast method, and uses a machine learning algorithm to detect footprints. However, despite several cell types being analysed and published, a software implementation was not made available. Another machine learning algorithm, CENTIPEDE¹⁰² can predict occurrences of motifs in DNaseI data. The limitation is the dependency on a list of motifs identified by ChIP-Seq in order to train the algorithm, in order to find the motifs in the data. The method is not able to make de novo predictions, or identify footprints at arbitrary genome locations. CENTIPEDE is available but requires pre-processing of the data in non-standard formats.

The algorithm used by ENCODE ³³ to identify all the footprints in human cell lines is based on a footprints occupancy score (FOS) which calculates the ratio of DNase-Seq tags within a binding site relative to those directly outside. This algorithm is not available, but all the footprints have been identified and released as part of ENCODE and can thus be visualised in the browser. This approach does not link the identified regions with any predictions of TFs from ChIP-Seq data, leaving that aspect to the user.

The most recent method for digital footprint calling is pyDNase, based on the so called Wellington algorithm ¹⁰³. This method accurately identifies depleted regions by assessing the difference between footprint shoulders and protection area while taking into account strand imbalance in the way the DNaseI cuts DNA. The algorithm identifies digital footprints in the data at a genome-wide level, without requiring any prior knowledge of motifs.

As it has been reported to be the best performing digital footprint caller ^{103,104} we tested pyDNase, with standard parameters DNase-Seq from erythroid cells. We ran the algorithm across 27465 HS regions and detected 78087 footprints. If we assume there are anywhere between 1 and 5-6 proteins directly bound in one hypersensitive site, forming footprints, then this is within an expected range. We decided to use pyDNase for the first assessment of digital footprints within regulatory regions in the different contexts analysed in this thesis.

(b) Digital genomic footprinting using DNase-Seq

DNase-Seq digital footprinting is based on the principle that, within a HS, bound proteins will protect the DNA from DNaseI digestion. If sequenced at sufficient depth, the frequency of DNaseI cutting can be used to identify the regions where a TF is directly bound to DNA. DNase digital footprints were visualised as described above using an in-house footprinting pipeline.

In summary, the frequency of enzyme cuts are assessed by quantifying the number of fragment ends within a HS. The frequency of cuts is displayed against the genome and results in a footprinting plot in which protein binding sites have

low frequency of cuts (Figure 25A). In the data, this translates in resolving TF binding sites within a hypersensitive site. An example can be seen in Figure 25B, where the footprint shows the binding sites of GATA1 within one of the enhancer HS of α -globin, R1.

(c) ATAC digital footprints have lower resolution than DNase digital footprints

Using the same analysis principles, we generated ATAC-Seq footprints and compared them with DNase-Seq footprints at the α -globin locus (Figure 26). At the enhancers and promoters, the ATAC-Seq footprints do not resolve binding sites to the same resolution observed in DNase-Seq footprints. Surprisingly, CTCF sites analysed within the locus show the same pattern in both data sets. In trying to identify the cause of the differences in footprinting we tested two variables: 1) sequencing depth and 2) enzyme processivity

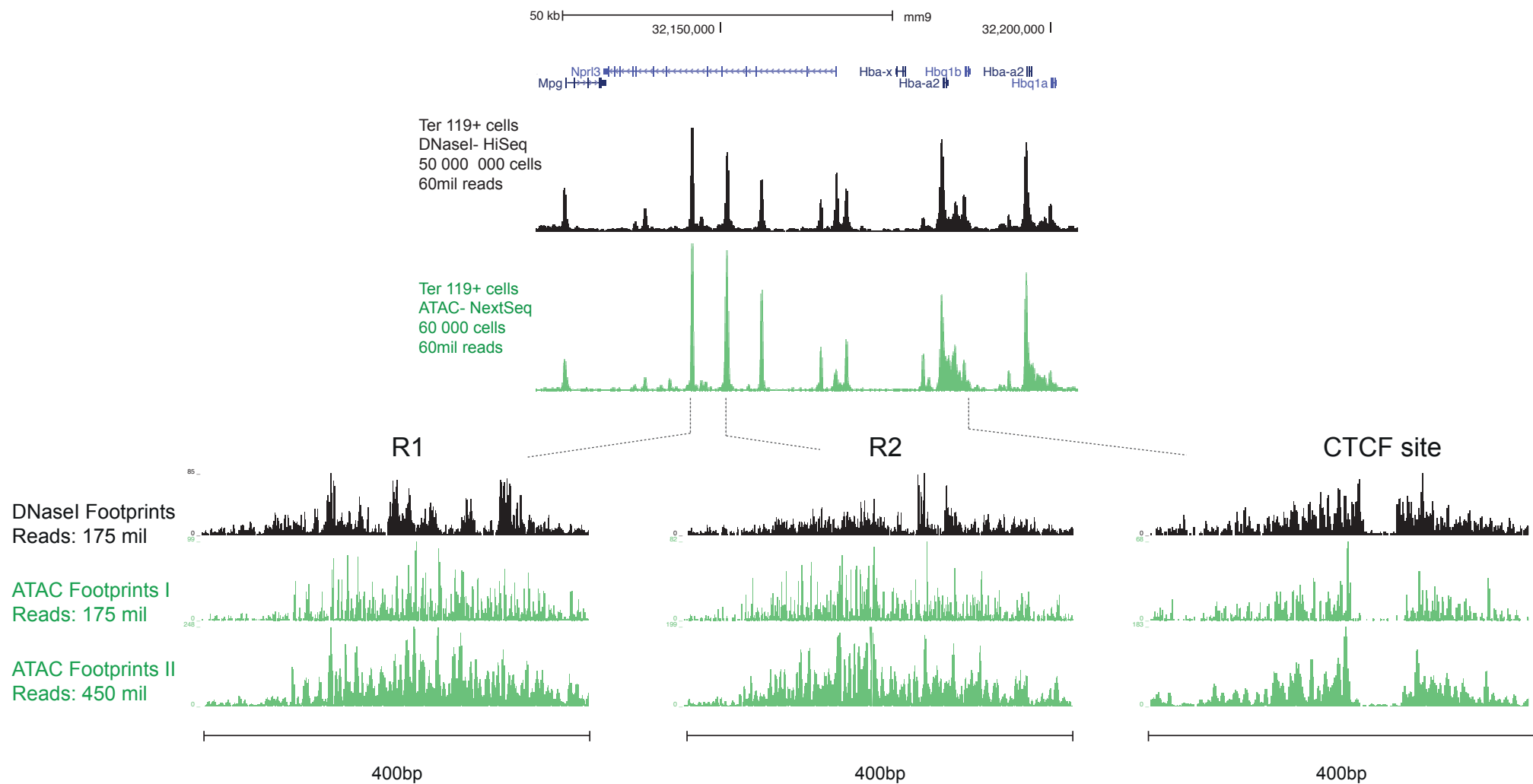


Figure 26. Comparison of digital footprints in DNase-Seq and ATAC-Seq. High-resolution analysis of the main regulatory enhancers (R1, R2) and the θ -globin CTCF site of the α -globin cluster (top). DNase-Seq footprints are shown in black. ATAC-Seq footprints generated at two different sequencing depths are shown in green.

Sequencing depth does not appear to make a significant difference. We generated a sample with higher sequencing depth, but did not observe a major improvement in terms of footprint resolution in the ATAC-Seq sample.

In order to assess enzyme processivity and exclude the frequency of cutting of the transposase as a potential cause, we quantified the frequency of cutting in a 400bp region over the same hypersensitive sites in DNase-Seq and ATAC-Seq (Figure 27). Unexpectedly, the modified Tn5 (transposase in ATAC-Seq) generates higher frequency of cuts than DNaseI, in the same window. Thus, it is not the lack of cuts, but maybe the location of the cuts within a HS region affected by enzyme mechanism.

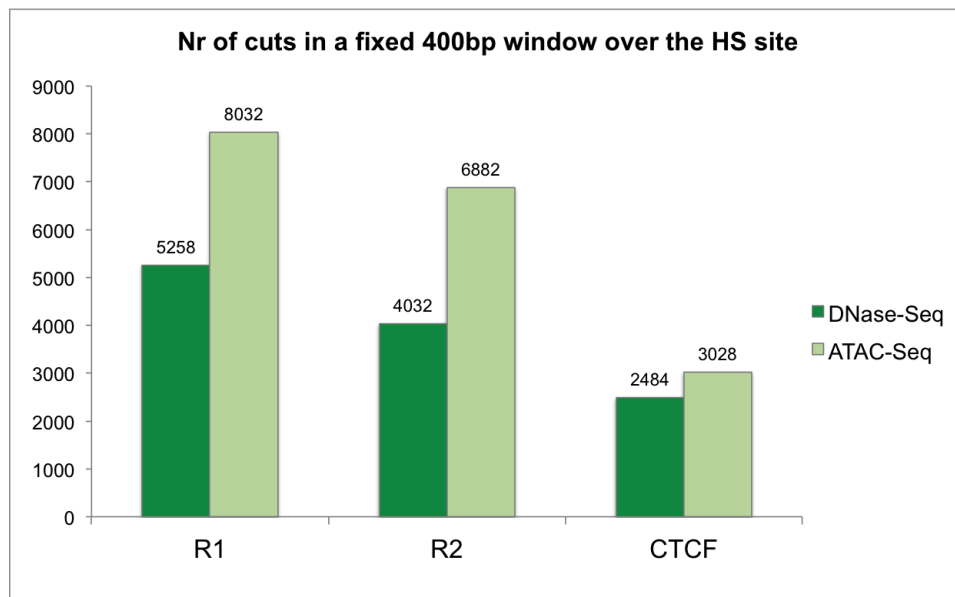


Figure 27. Enzymes processivity derived from number of cuts in a 400bp window. Assessment of DNaseI and Tn5 processivity. Cuts generated in a window of 400bp from DNase-Seq (dark green) and ATAC-Seq (light green) were quantified at the same sites: R1, R2 and θ -globin CTCF.

(d) Conclusion II

To summarise, ATAC-Seq can reliably and more efficiently isolate open chromatin regions resulting in overall higher HS signal than DNase-Seq. However, in terms of high-resolution analysis of ATAC-Seq, the signal is poor and digital footprints cannot be detected. This might be due to multiple causes (See Discussion, Section I), including input material complexity or the mechanism employed by the transposase to fragment the open-chromatin regions. Thus, for generation of digital footprints, DNase-Seq is still the method of choice resulting in high quality footprints.

Having optimised and setup the necessary methods, we are now in a position to analyse changes in HS regions, at low and high resolution, in development, differentiation and disease.

Chapter II. Analysis of *cis*-regulatory elements in erythroid development

2.01 Introduction

Mammalian primitive and definitive erythroid cells are characterised by distinct origins and are present in the mouse embryo at different developmental stages (primitive from E7.25 to E12.5, definitive from E8.25 -E17.5). They present with distinct morphological features (different size, haemoglobin content) (Figure 28) and most remarkably, express different globin genes from both alpha and beta clusters (see Introduction, Primitive erythropoiesis).

As the embryo develops, the embryonic ζ - and ϵ/γ , and foetal β H1- globins switch to adult α - and β -globin expression (See Introduction, Figure 5). Extensive studies of this phenomenon termed “switching” were driven not only by interests in the transcriptional machinery behind this phenomenon but more importantly by the potential therapeutic implications for treating hemoglobinopathies ^{79,105-107}. Despite the immense interest in the topic, our understanding of this phenomenon is still limited.

For the purpose of this thesis, we will focus on ζ -globin expression in primitive cells, the embryonic gene in the α -globin cluster, and its transcriptional regulation as compared to the α -globin regulation.

The mouse ζ -globin promoter is located in the α -globin cluster on chromosome 11, upstream of the adult globin promoters. Conservation analysis⁵¹ showed that the ζ -globin promoter is highly conserved across 22 species of vertebrates, including human and mouse.

Distal regulation of the ζ -globin promoter has been characterised in the human α -globin cluster. In vitro studies showed that the major regulatory element of the adult α -globin (HS-40) can derepress the ζ -globin promoter¹⁰⁸, indicating that the enhancers of the adult α -globin genes may be involved in the regulation of the embryonic/fetal genes in vivo. Whether all the enhancers of the adult α -globin LCR are active in vivo in primitive erythroid cells and involved in the regulation of the embryonic ζ -globin promoter remains to be addressed. Data from the lab on the chromatin interactions in primitive cells suggests that the ζ -globin promoter is interacting with the LCR in the mouse. However, our understanding of how the regulation is coordinated between the embryonic and adult genes is still limited. Repression of ζ -globin in definitive erythroid cells is thought to be achieved by the same molecular mechanism observed at the embryonic beta globins, involving Bcl11A and Sox6^{105,109}.

To add insight into the mechanism of embryonic globins regulations, we are examining regulatory sequences accessible to and bound by TFs in both primitive and definitive erythroid cells.

In this chapter, we first focus on the comparison of the HS regions at low and high resolution, at the α -globin locus, in primitive and definitive erythroid cells. We examine the enhancer sequences, α -globin promoter and the ζ -globin promoter. Second we compare, at a genome-wide level, the HS regions in primitive and definitive erythroid cells and relate significant changes in hypersensitivity to published differential gene expression in both cell types.

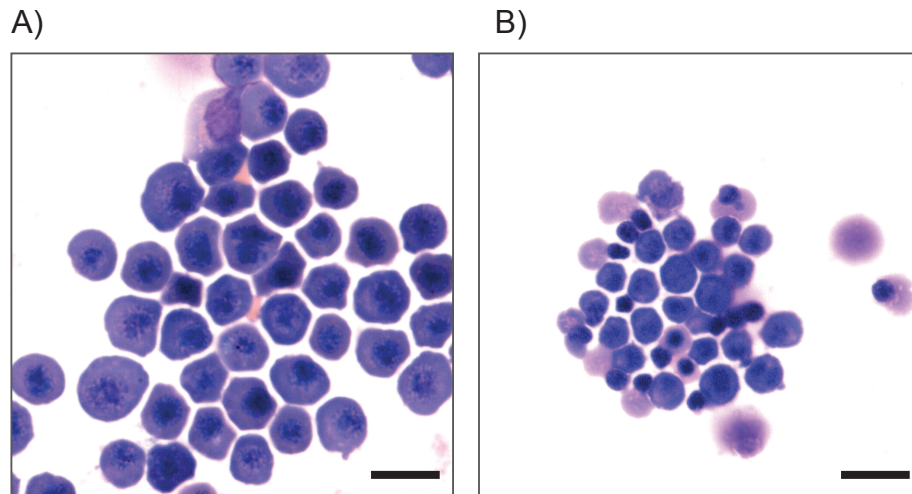


Figure 28. Morphology of erythroid progenitors from different developmental stages. A) Primitive erythroid progenitors, from circulating blood, stage E10.5. B) Definitive erythroid progenitors from E12.5 foetal liver, cultured for 30 hours in high Epo media to induce terminal differentiation. Cytospun cells stained with May-Grünwald Giemsa stain. Scale bar 20 µm. Representative photos from more than three independent experiments.

2.02 Chromatin accessibility at the α -globin cluster in primitive and definitive erythroid cells

(a) Open chromatin assays reveal a similar hypersensitive landscape in primitive and definitive cells

In order to investigate sequences accessible and bound by TF binding at the α -globin locus and the ζ -globin promoter, we performed open chromatin assays (DNase-Seq, ATAC-Seq) in primitive and definitive erythroid cells (Figure 29).

Circulating primitive cells were isolated from E10.5 mouse embryos. These experiments were part of a collaboration with Dr Duantida Songdej, a member of our unit. A stringent morphological and molecular characterisation of the primitive cells was performed by Dr Duantida Songdej and will be discussed more extensively in her thesis. Our collaboration was restricted to comparing the chromatin landscape and potential TF bound sequences at the α -globin cluster in primitive and definitive erythroid cells. Definitive Ter119⁺ erythroid cells were obtained from spleen of APH-treated mice and foetal liver tissues (Materials and Methods).

Our data presents, for the first time, regions of open chromatin in primitive erythroid cells.

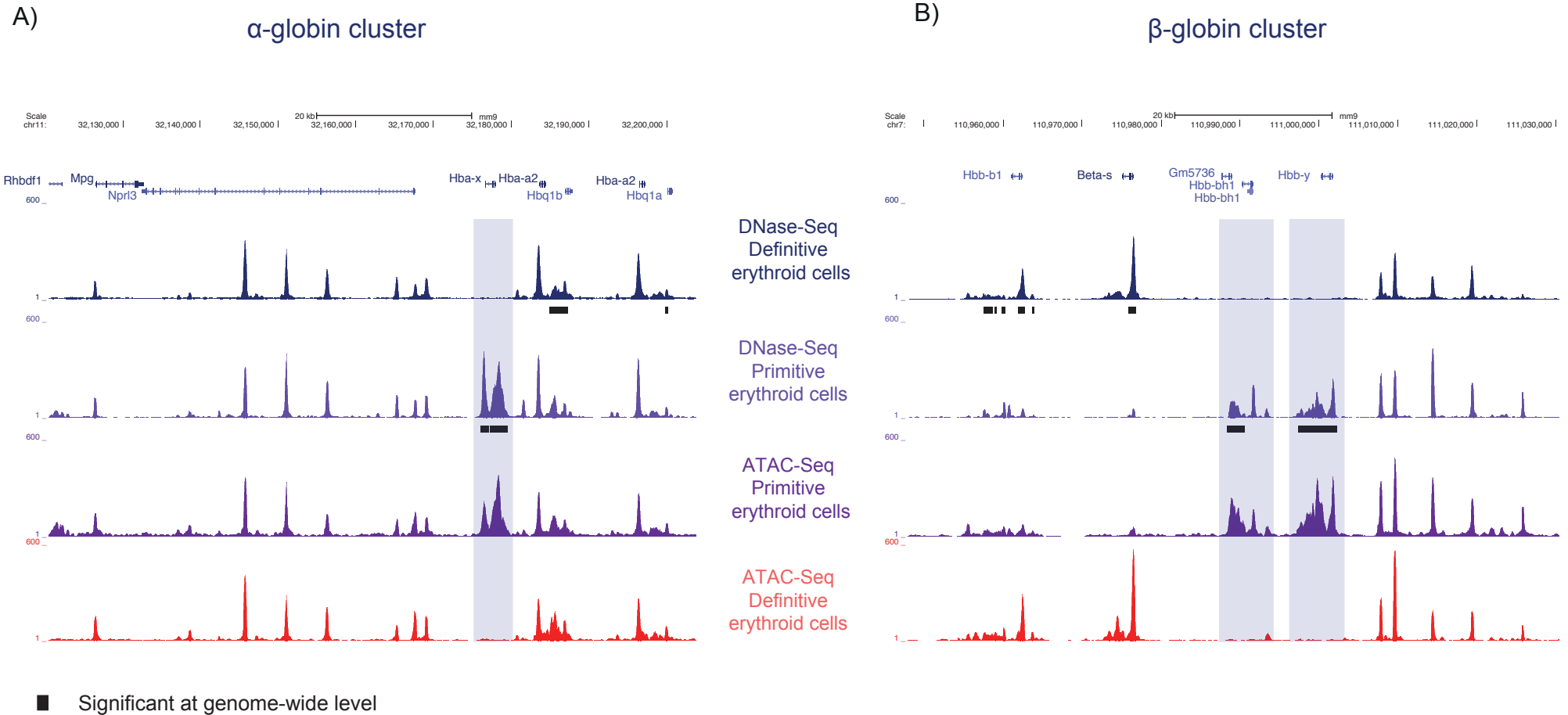


Figure 29. Open-chromatin landscape at the globin loci in mouse definitive and primitive cells. The HS regions at the α -globin cluster (A) and β -globin cluster (B). Peaks represent areas of open-chromatin enrichment and correspond to regulatory elements previously described. Highlighted are the zeta-globin promoter (left) and the fetal/emrbyonic promoters (right) marked by hypersensitivity in primitive but not in definitive cells.

As expected, the embryonic ζ -globin promoter is hypersensitive in primitive and not definitive erythroid cells, suggesting binding of TFs necessary for the transcriptional activity of this embryonic gene (Figure 29). This acquired hypersensitivity corresponds to statistically significant enrichment not only at the locus but also at the genome-wide level (discussed below). Interestingly, an additional peak is observed at the 3' end of the ζ -globin gene. This statistically significant enrichment at the end of the gene is unexpected and, as yet, of unknown significance.

As for the *cis*-elements in the cluster, all five enhancers are present in both cell types including the major regulatory elements of the adult α -globin genes, R1 and R2. The adult α -globin promoters are hypersensitive in both cell types, though they show significant increase in hypersensitivity between the gene ends and the neighbouring Hbq promoters in definitive cells. This increase in hypersensitivity is associated with the increase in transcription as definitive cells express adult α -globin at significantly higher levels than primitive cells¹¹⁰.

At the β -globin locus, the embryonic ε - and foetal β H1 promoters are hypersensitive only in the primitive erythroid. The adult β -globins also show an increase in hypersensitivity associated with the transcriptional activity of these adult genes in definitive cells. These observations are in line with the literature on the characteristic expression of embryonic/fetal globins in primitive cells.

Analysis of open-chromatin regions highlights the hypersensitive regions at a resolution that does not reveal sufficient information about the sites within the regulatory regions that are bound by TFs or the nature of the TFs recruited to these regions. In order to increase the resolution of the hypersensitive sites, we generated genome-wide DNaseI footprints (Materials and Methods).

(b) - Digital DNase-Seq footprinting of the adult α -globin promoters and enhancers shows no significant difference between primitive and definitive erythroid cells

High-resolution analysis at the α -globin cluster reveals a similar pattern of digital footprints between primitive and definitive cells at the enhancers and adult α -globin promoters. This suggests that the same regulatory regions are directly bound by TFs in the two cell types (Figure 30). Those sequences were further examined in the UCSC browser using conservation analysis and by searching for erythroid-specific TF motifs.

The previously described TF binding sites, both highly conserved and mouse specific⁵¹, underlie digital footprints at the major regulatory elements and promoter. In R1 we can identify the two conserved GATA1 sites, as well as two mouse specific sites; in R2 three AP1 motifs, for NF-E2 binding are visible (two conserved, one mouse specific)^{42,53}. At the α -globin promoter, we can identify the GATA1 site, and the conserved CCAAT binding site⁵¹.

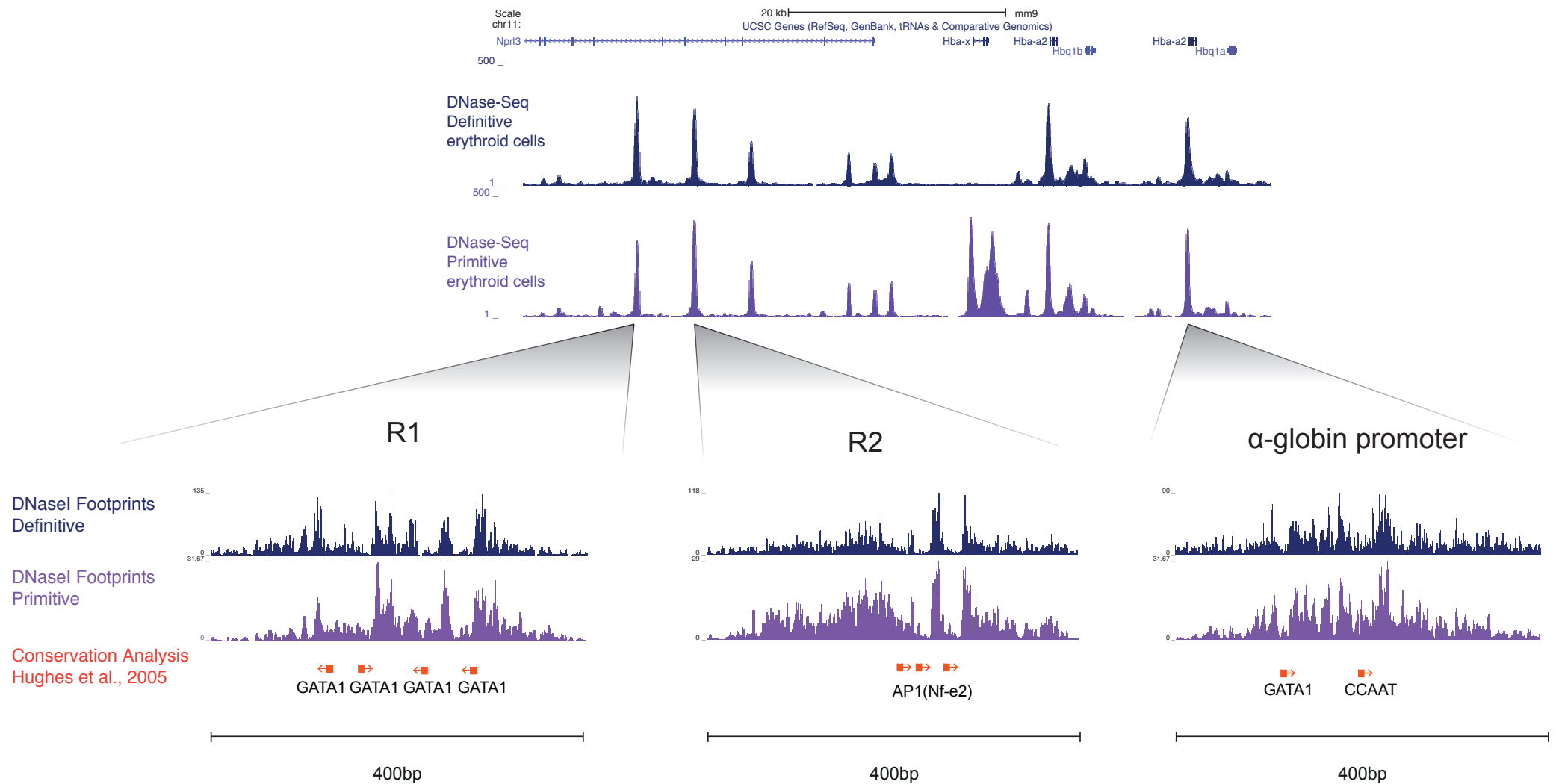


Figure 30. DNase digital footprints at the α -globin cluster. UCSC tracks showing the HS regions in primitive and definitive regions at the α -globin cluster. High-resolution analysis of DNase-Seq at the major regulatory elements R1, R2 and the α -globin promoter reveal the digital footprints within the HS regions. The previously conserved binding sites are marked by red boxes.

To summarise, following high-resolution analysis, the same binding sites appear occupied at the five enhancers and the promoter of the α -globin cluster. This suggests that either (1) the direct binding at these elements is the same in the two same cell types; thus the differences in transcriptional regulation (expression of ζ -globin) are achieved at the level of protein-protein interactions and not at level of direct DNA binding; or (2) the regulatory regions are bound by different TFs from the same families and recognising the same motifs. This could be clarified by ChIP-Seq experiments for TFs such as GATA1 or NF-E2.

(c) Digital DNaseI footprinting of the ζ -globin promoter confirms some of the observations made by in vitro DNaseI footprinting

We next analysed the DNase-Seq digital footprints at the ζ -globin promoter. By visual assessment, two footprints stand out and could be identified as the conserved TATA-box and E-box motif. No other footprints could be identified.

With the aim of computationally detecting the footprints ζ -globin promoter we used the pyDNase footprint caller (Chapter 1- Footprinting detection, Materials and Methods). pyDNase detects all the areas with low frequency of cutting, and identifies 6 digital footprints which partially overlap with vertebrate conserved regions (Figure 31). These detected sequences confirm footprints previously identified in vitro in the human ζ -globin promoter^{111,112}.

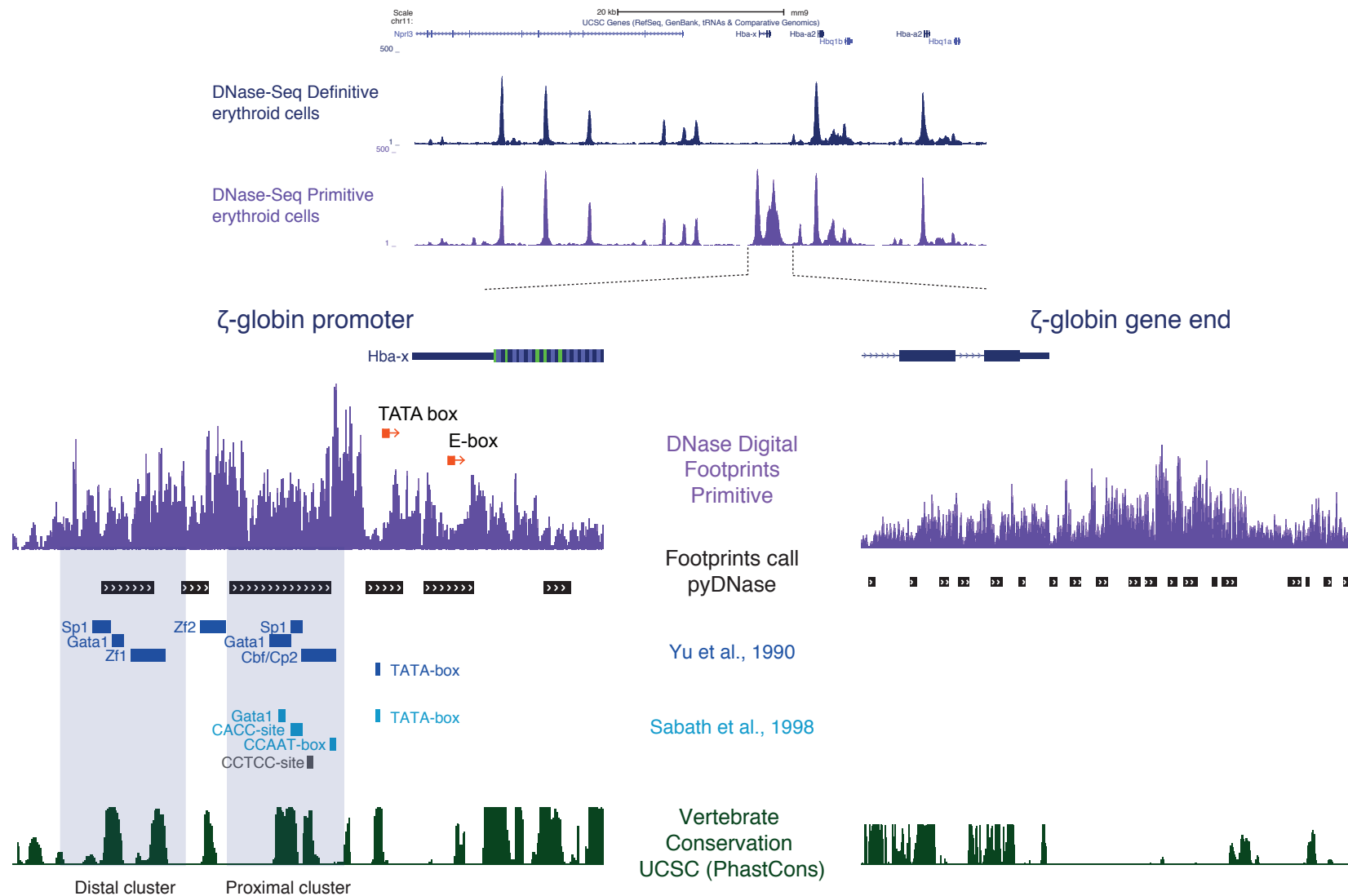


Figure 31. Comparison of digital footprinting at the ζ -globin region with previous studies. A) Digital footprints at ζ -globin promoter and the ζ -globin gene. Significant footprints are detected with pyDNase and compared to results from in vitro studies (Yu et al., 1990), in vivo studies and site directed mutagenesis (Sabath et al., 1998) and vertebrate conservation (UCSC). The HS regions at the α -globin cluster are shown on top.

Protein binding at the human ζ -globin promoter has been analysed and reported by two studies, Yu et al., 1990 and Sabath et al., 1998. Considering the lack of mouse- specific studies, but more importantly, the high degree of conservation between the mouse and human ζ -globin promoter, we will compare our findings from the murine locus with the ones previously reported from human cells.

Yu et al. performed DNaseI footprinting using gel mobility shift assay with human erythroid K562 and non-erythroid HeLa cell extracts. They identify two clusters of binding motifs in the ζ -globin proximal promoter (Figure 31), containing amongst others, GATA1 sites, CACC sites thought to be Sp1 binding sites, and CCAT sites. Those sites are thought to potentially play a role in the transcriptional activation of the ζ -globin gene, though activity was not assessed in this study¹¹¹.

In contrast, Sabath et al., use site-directed mutagenesis and luciferase reporter assay in erythroid K562 cells and ζ -globin non-expressing OCIM1 cells to assess the sites' contribution to ζ -globin activity¹¹². Sabah et al., confirm some of the previous results: the promoter-proximal sites are important for ζ -globin activity. However, they do not find SP1 site important for activity, and cannot identify binding of Sp factors to that site. Thus, in terms of proteins, except for GATA1, there is little consensus on which TFs are bound at the ζ -globin promoter from the in vitro studies.

We confirm by digital DNaseI footprinting, in primary primitive cells, that the sites previously reported are bound by TFs. The two clusters of binding sites are identified as statistically significant footprints, and the regulatory sequences

bound are confirmed. The current analysis of DNaseI footprints does not provide further information in regards to the proteins bound. The task of identifying the full complement of proteins bound to these sites requires additional experimental methods (proteomics), is beyond the scope of this thesis and remains to be addressed.

(d) Hypersensitive site at 3' end of the ζ -globin gene is of unknown significance

There are no previous studies reporting the hypersensitivity of the 3' end of ζ -globin gene. This hypersensitive site is however reproducible in both open chromatin assays, across replicates and it is a statistically significant enriched site in primitive when compared to definitive cells. There are 21 footprints identified by pyDNase, but none of these regions could be further characterised.

This 3' end hypersensitivity can be observed at the β -embryonic and fetal genes as well (Figure 29) and is reminiscent of the adult globins hypersensitivity pattern. In the literature, gene 3'-end hypersensitivity has been scarcely reported, to have either a structural role, or a regulatory role.

In the study of the human apolipoprotein B, the 3'-end hypersensitivity was thought to be involved in anchoring the gene to the nuclear lamina¹¹³. Association of chromosomal domains with the nuclear lamina has been previously reported to be linked with gene silencing, though the cause-effect relationship of the mechanism is still unclear¹¹⁴, but this might be the case for the silencing of the ζ -globin promoter. This could be confirmed by imaging studies. In terms of role in

regulation, a similar site was reported to have a role of tissue specific enhancer in T cells, at the CD3-delta gene 3'end¹¹⁵. It might be that this region acts as a primitive specific enhancer. Initially, this could be verified by analysing chromatin marks such as H3K4me1 in primitive cells.

One of the main differences between primitive and definitive cells is the transcriptional output of the globin genes clusters. We show that this difference is also reflected in the levels of hypersensitivity of the embryonic genes promoters. The digital footprints show that at the level of protein-DNA interactions, the different regulation of the α -globin cluster in primitive cells is achieved through a combination of TFs binding to primitive specific regions, but also sites common with definitive regions.

In order to identify the regions where TF binding is cell-specific, we performed qualitative and quantitative genome-wide analysis of the HS regions in primitive and definitive cells.

2.03 Genome-wide analysis

(a) The differences in hypersensitivity are mainly found at regulatory regions

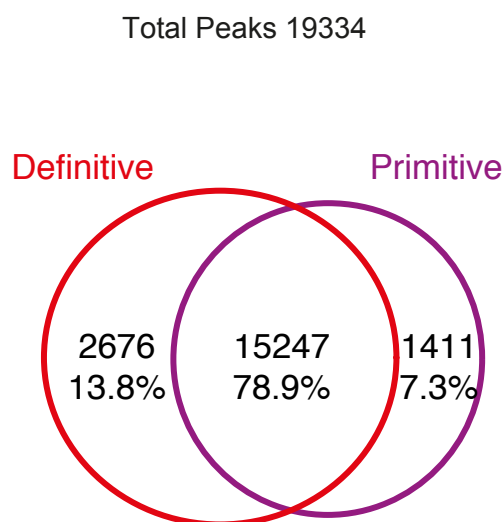
Main regulators of developmental changes such as switching from the primitive to the definitive erythroid cells are *cis*-regulatory regions (mainly enhancers), that present as platforms of regulatory sequences for binding of tissue specific TFs. In order to study TF binding at regulatory sequences at both enhancers and promoters in primitive and definitive erythroid cells, we performed genome-wide analysis of the open-chromatin landscape. The differences were assessed by studying the significance of differential enrichment in hypersensitivity.

In order to perform differential analysis, three biological replicates are required. Due to the low input material required for ATAC-Seq (10^5 cells instead of 10^7 cells required for DNase-Seq) (Materials and Methods/ Chapter 1), this was the assay of choice for generating the data sets necessary for the statistical analysis, as the primary cell material available from primitive erythroid cells is limited.

The enriched regions (peaks) were identified using the peak calling algorithm described in the Materials and Methods section. After filtering the low reads regions, we identified 16658 peaks in primitive cells and 17923 peaks in foetal definitive erythroid cells. We performed differential analysis as described in (Materials and Methods/Chapter 1) and found 1411 (7.3%) peaks significantly enriched in primitive cells, and 2676 (13.8%) significantly enriched in foetal

definitive cells. 78.9% of sites do not show significant enrichment changes between the two cell types (Figure 32).

A)



B)

Differential Peaks: Definitive (2676) vs Primitive (1411)
(4087 peaks, $\log_2$ fold change >2 , FDR < 0.010)

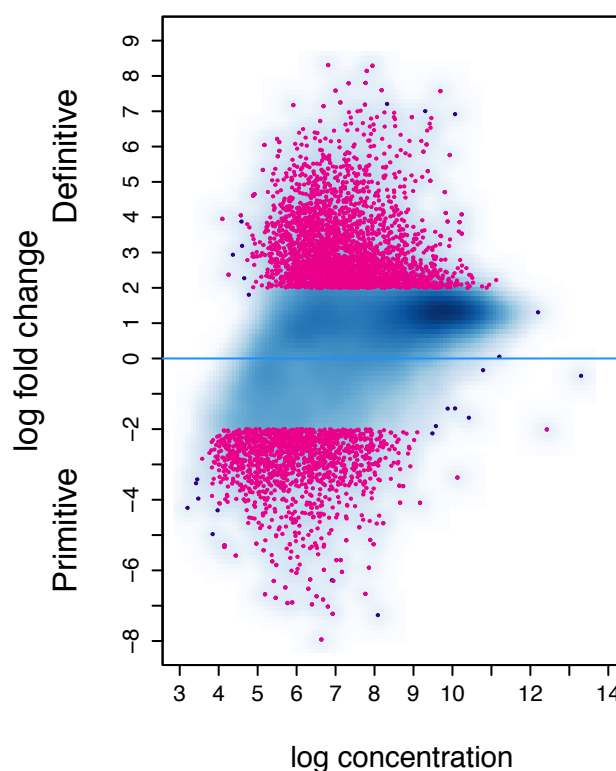


Figure 32. Overlap and differential peaks in Primitive and Definitive ATAC-Seq. A) Venn diagram illustrating the proportion of peaks, out of total, that are differentially enriched ($\log_2$ fold change >2 and FDR < 0.01) in definitive and primitive erythroid cells. B) Fold changes in read density ($\log_2$) between two samples (y-axis) are plotted against the average of read density (x-axis) at the same peaks. Differential peaks in Definitive cells will have a positive log fold change, while differential peaks in Primitive cells will have a negative fold change. Differential analysis with three independent biological replicates.

Thus, the majority of HSs do not differ between the two primitive and definitive erythroid cells. In order to characterise the cell-specific regions which we hypothesise are enhancers driving cell type specific differences, we annotated the differential peaks.

(b) Annotation of differential regions

We characterised the total and differential peaks based on chromatin marks available only from the definitive erythroid population. We annotated 43% promoters, 29% enhancers and other genomic elements in the total peaks set (Figure 33). When annotating the differential peaks, we observe that the largest proportion (66%) of definitive-only peaks are enhancers and 14% are promoters. Interestingly, the majority of primitive peaks do not correspond to definitive erythroid chromatin marks, suggesting that the location of the *cis*-regulatory elements differs in primitive and definitive erythroid populations. This is interesting as it confirms that HSs are lost and gained as cells progress through development and differentiation³⁷. This is also not unexpected considering primitive and definitive cells are different cell types, with specific chromatin marks pattern and thus in order to properly annotate the primitive regions, ChIP-Seq data from this cell type is necessary.

However, these primitive peaks not previously characterised can be annotated based on genomic location only, using UCSC and RefSeq annotations. This is less informative than histone marks annotation but shows that more than 60% of the peaks are intergenic and intronic, characteristic features of enhancer elements (Figure 34).

Peaks Annotation

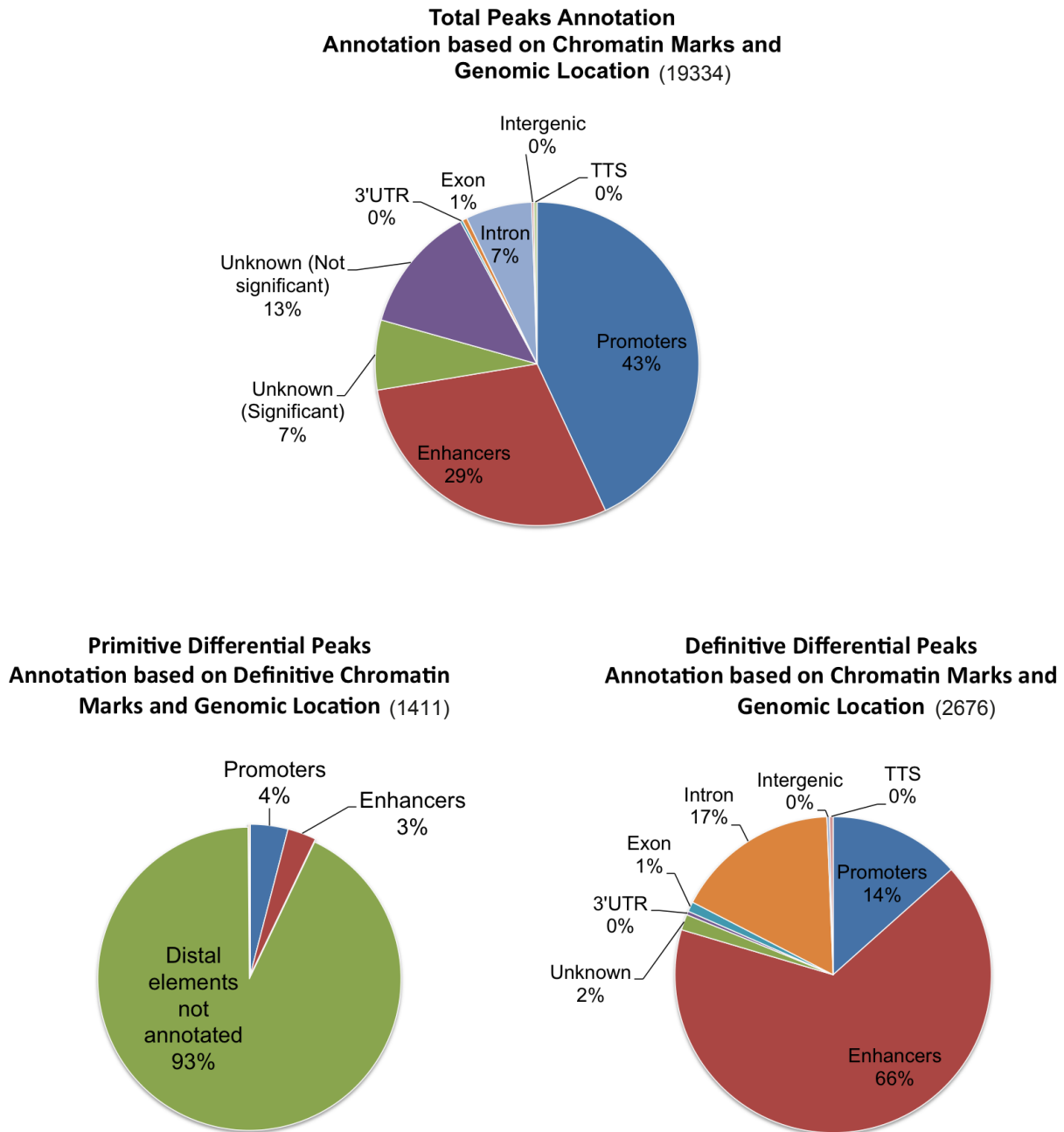
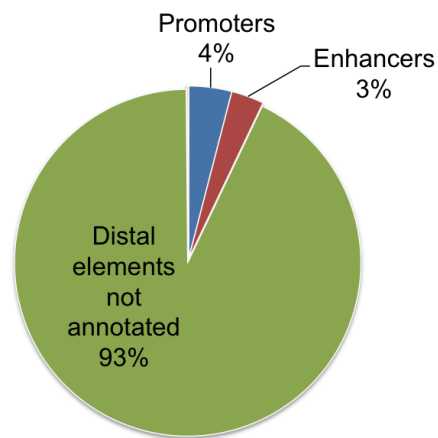


Figure 33. Peaks Annotation based on genomic location and chromatin marks H3K4me1/H3K4me3 ratio. Piecharts showing the annotation of total peaks analysed, as well as of the differentially enriched peaks in Primitive and Definitive cells. Enhancer and promoter elements were annotated based on H3K4me1/H3K4me3 ratio in Definitive cells, as well as on genomic location from publicly available annotations. All the other elements were classified based on genomic location. The unknown fraction from the Primitive cells could not be reliably identified based on chromatin marks ratio.

**Primitive Differential Peaks
Annotation based on Definitive Chromatin
Marks and Genomic Location**



**Primitive Differential Peaks
Annotation based on Genomic Location**

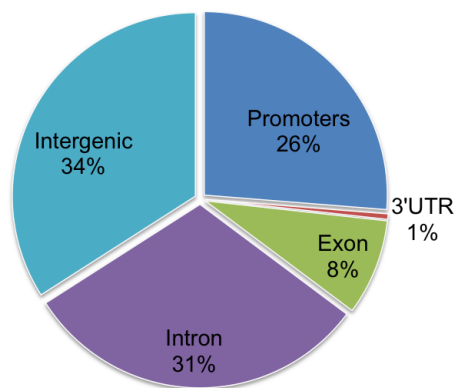


Figure 34. Annotation of unknown peaks based on genomic location. The regions that could not be identified based on the enhancer/promoter call from the Definitive cells were annotated based on genomic location only, using UCSC and RefSeq annotations.

(c) Correlation of differential open chromatin regions with differential gene expression

Differential gene expression profiles between primitive and definitive erythroid cells is available¹¹⁶ and could be used with our differential open chromatin dataset to elucidate aspects of the molecular mechanisms (TF bound sequences) that drive the transcriptional programmes between the two cell types

We investigated the correlation between our identified differential hypersensitive sites and published cell specific gene expression by analysing a selection of differential genes in our data, in both primitive and definitive erythroid cells.

Kingsley et al. highlight Pbx1, Foxh1 and Arid3a as primitive erythroid specific TFs. We examined regions of hypersensitivity in these gene loci that would potentially explain the primitive-specific expression pattern. Although Pbx1 promoter sequence is not significantly enriched in primitive erythroid cells, intragenic elements, most likely enhancers, are hypersensitive only in primitive and not definitive (Figure 35 A). As for Foxh1, our data detects an increase in hypersensitivity at the promoter in primitive erythroid cells though not statistically significant. Arid3a, shows a significant increase in hypersensitivity at the promoter of an isoform, which is present only in primitive cells.

In definitive erythroid cells, the most significant gene expression changes are observed for Sox6, Nr3c1, Cebpa, and Irf9. Supported by data available on chromatin marks from definitive cells, the changes observed in the hypersensitivity at these loci could be correlated with enhancers and promoters. Sox6 shows significant differential enrichment at enhancers, and the promoter of

an isoform (Figure 35B). Nr3ca has a significant increase of hypersensitivity at the promoter. Interestingly, even though differentially expressed, Cebpa does have a hypersensitive promoter in either cell types, whereas Irf9 has very similar levels of hypersensitivity in both primitive and definitive cells.

In addition to the TFs selected by Kingsley et al., we also investigated the Bcl11a gene, a known repressor of the fetal globin genes¹⁰⁵. Both the Bcl11a promoter and intragenic enhancers are only hypersensitive in definitive erythroid cells, as expected.

Differential hypersensitivity at enhancers and promoters of these genes suggest that some differences in gene expression could be attributed to tissue specific distal regulation. However, the discrepancies between the hypersensitive regions and gene expression differentials in some examples confirm that beyond the simple case of associating a hypersensitive promoter/enhancer with gene expression, and lack of hypersensitive peak with absence of expression, there is a more complex level of regulation. This involves processes such as the rate of transcription, accumulation of transcripts or decay rates, and could be further complicated by recruitment of cofactors and protein complexes that modulate transcription differently in different biological contexts.

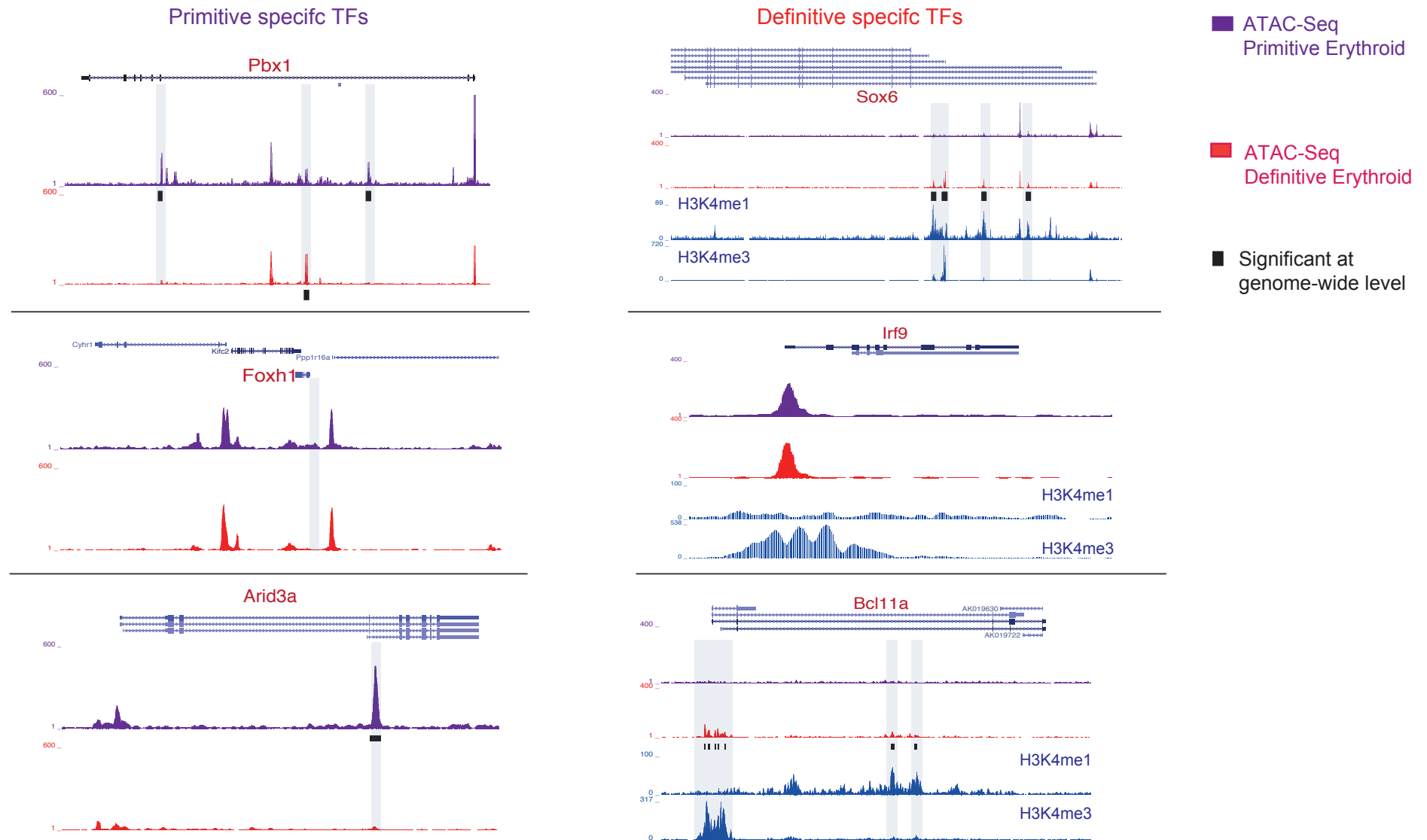


Figure 35. The hypersensitivity pattern at cell-specific expressed TFs. Selected examples of significantly expressed TFs (Kingsley et al., 2013) in primitive and definitive cells. UCSC tracks show HS regions as determined by ATAC-Seq in primitive (purple) and definitive (red) erythroid cells. For definitive cells, chromatin marks are available and displayed in blue. RefSeq genes are shown on top.

2.04 Conclusion

Even though primitive and definitive erythroid cells share morphological features and the physiological function of producing red blood cells, they differ in size, globin expression and global gene expression. The differences highlight the fact that distinct cell types can fulfil the same biological function, at different times in development, through gene regulation.

To gain insights into the different mechanisms of regulation, we analyse HS regions at two different levels: single locus (α -globin) and genome-wide.

At a locus-specific level, we show for the first time the open-chromatin landscape in primitive cells. Through quantitative analysis, we confirm that transcriptional activity of the embryonic genes in primitive cells is reflected by statistically significant HS enrichment.

Analysis of the other regulatory elements shows similar levels of HS, making the regulation of the ζ -globin difficult to untangle from the one of the α -globin. We explore this further, by high-resolution analysis. Digital DNase-Seq footprints analysis reveals that the TF occupancy of the enhancers and adult α -globin promoters is strikingly similar in primitive and definitive cells. This suggests that the differential transcriptional output is not achieved through different direct binding of the enhancers/promoters, but probably at an indirect level.

We must keep in mind, however, that based on the footprinting methodology, we cannot draw a conclusion about the exact TFs bound. There is a possibility that different TFs belonging to the same family (i.e. GATA, KLF) are recognising these

regions. These observations could be confirmed by ChIP-Seq in primitive cells, though these experiments are still problematic due to the low amounts of primary material.

At a genome-wide level, we show through quantitative analysis, that ~10% of regulatory elements are enriched in each cell type. These statistically significant regions are most probably enhancers in primitive and confirmed enhancers in definitive, reinforcing the dogma that distal elements drive cell specific differences.

The direct binding at these differential regions is of extreme interest to us and will be addressed in Results Chapter 4.

Chapter III. Analysis of *cis*-regulatory elements in erythroid differentiation

(a) Introduction

In order to identify changes in direct TF binding at regulatory sequences that drive differentiation in general and activation of α -globin in particular, we examined the differences in HSs in the mouse erythropoiesis. Open-chromatin was assessed genome-wide and at the α -globin enhancers, as the cells transition from early progenitors to differentiated erythroid cells.

We know from previous studies that the activation of α -globin expression is achieved through a step-wise process of chromatin acetylation, HS formation, tissue specific TF binding, and PolII recruitment at the upstream regulatory regions and the promoters^{52,53,81}.

The presence of HS sites was assessed at the α -globin locus as differentiation progressed. Main players in the early setup of the locus were identified to be the GATA factors 1 and 2, NF-E2 and SCL⁵³. These factors are thought to be recruited to the conserved binding sites within the HSs⁵¹. For the activation of the promoter, Sp/X-KLF factors were shown to be recruited prior to activation, concomitantly with the pre-initiation complex⁸¹. Those studies were limited by the technology available at the time, which affected the resolution they can achieve and the number of loci that can be assessed. In addition, the amount of input material necessary for the assays made the use of primary cells prohibitive.

Several questions remain unanswered regarding the mechanism of TF binding at the regulatory sequence: Does usage of the binding sites within a HS differ between different stages of erythropoiesis? How does this relate to chromatin marks and enhancer/promoter interactions and ultimately to transcriptional activity?

We set out to address some of these questions by analysing hypersensitivity at low and high resolution. More specifically, in this chapter, we aim to identify the changes in HS both at the α -globin locus and genome-wide in primary definitive erythroid cells. We quantitate the changes in open-chromatin in purified early progenitors and differentiated erythroid cells, to assess if the level of hypersensitivity is indicative of activity. We characterise histone modifications and chromatin interactions at the α -globin cluster, at the two stages of differentiation, and relate them to observed changes in open-chromatin. Next we generate footprints to study the changes at high resolution in TF binding at the model locus as the cells differentiate from early progenitors to mature cells. Finally, we perform genome-wide analysis to analyse the global differences in open-chromatin landscape between the two stages of differentiation.

3.02 Purification of primary erythroid populations corresponding to two different stages of erythropoiesis

(a) Isolation of early progenitors and differentiated erythroid cell populations

To address the molecular dynamics of differentiation, it is crucial that we isolate two distinct populations that mark the early and late stages of erythropoiesis (Figure 36).

Cell populations were isolated as described (Figure 6, Materials and Methods), using a magnetic column purification strategy based on the following immunophenotypes^{117,118}: Lin-/ckit+/CD71+/Ter119- cell population corresponding to early erythroid progenitors and Ter119+/CD71+ population corresponding to differentiated erythroid cells.

Those two populations were further characterised to confirm their nature using cellular staining, immunophenotyping, and the differentiation potential of the progenitor population in vitro. Most importantly, we verified that the expression of the α -globin genes reflect the stage of differentiation in these cells.

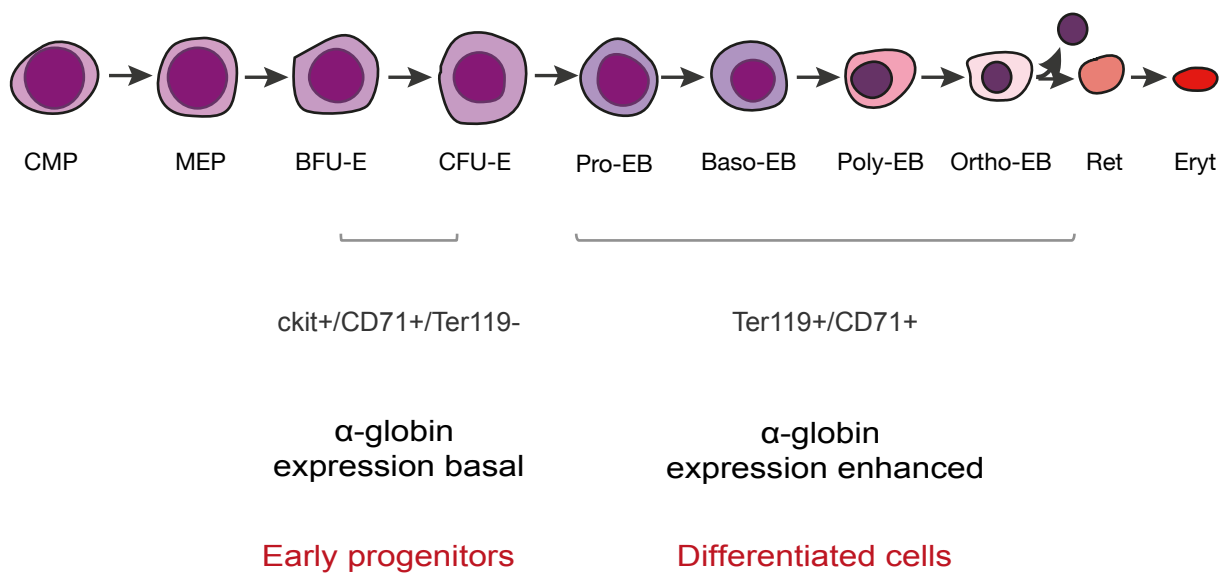


Figure 36. A schematic representation of erythropoiesis depicting cells from early progenitors to enucleated red cells. The isolation of early (α -globin expression basal) and differentiated (α -globin expression enhanced) cell populations is based on expression of the surface markers ckit, CD71 and Ter119, as indicated.

(b) Morphologic and immunophenotypic assessment of the two isolated populations by cellular staining

MGG staining was performed to assess the morphology of the cells (Materials and Methods) (Figure 37A). The isolated early erythroid population is morphologically homogeneous as expected for committed erythroid progenitors (BFU-E/CFU-E) and presenting as large cells with deep-blue cytoplasm and visible nuclei that occupy about 3/4 of the cells. The differentiated population represents a mix of maturing erythroid cells ranging from pro-erythroblasts to pyknotic red cells. They are generally smaller than the progenitors, with varying degrees of pink in the cytoplasm (indicative of haemoglobinisation) and compact nuclei⁷⁸.

FACS analysis confirmed the immunophenotypes of the magnetic column purified populations: the early Ter119- progenitors express high CD71 and high ckit (93.1% of the population), while the differentiated cells are marked by high CD71, high Ter119 (96.3% of the population), and low ckit (Figure 37B). The separation of the ckit and Ter119 expressing cells between the two stages was clear and reproducible (Figure 37C) and ensured the purity of the isolated populations.

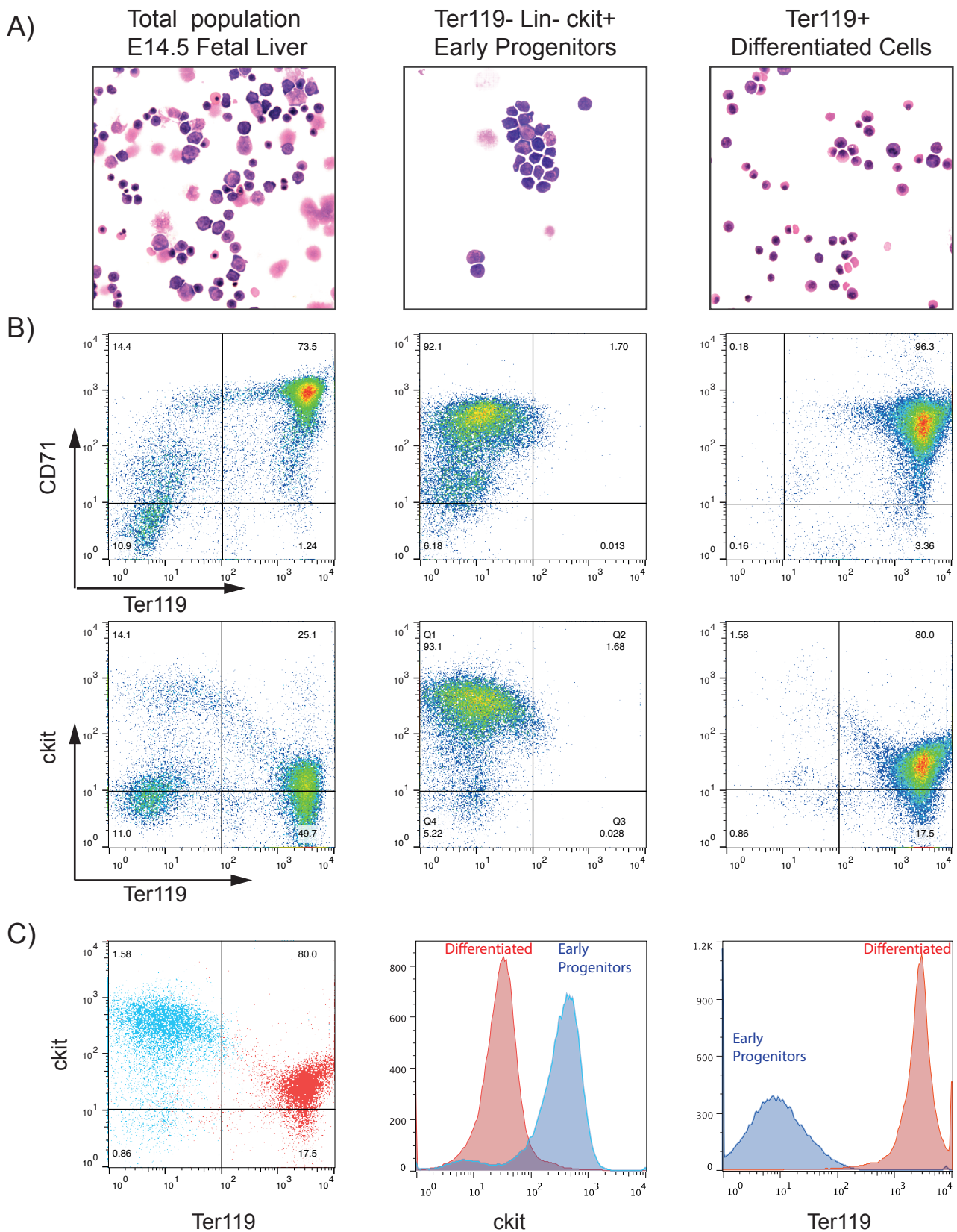


Figure 37. Characterisation of early progenitors and differentiated populations, isolated from E14.5 fetal liver. A) Giemsa- stained cytopspin preparations from total and selected populations. Early progenitors: Ter119-Lin1-ckit+; Differentiated cells: Ter119+, CD71+. B) FACS plots displaying the cell surface markers of each population. C) FACS plot and histogram confirming the separation of the ckit and Ter119 markers used for isolating the two populations.

(c) Developmental potential of the early progenitors

The developmental potential of the early progenitors isolated can be assessed by their ability to form colonies in semi-solid media. Upon plating the purified progenitors in methylcellulose supplemented by growth factors (Materials and methods), we observed formation of CFU-E (Colony Forming Unit-Erythroid) at day 4 and BFU-E (Bursts Forming Unit-Erythroid) at day 6 (Figure 38). We observed hemoglobinisation in both BFU-Es and CFU-Es. FACS analysis showed that the level of ckit expression was reduced from day 4 to day 6, while the level of Ter119 increased, confirming that the cells were differentiating normally.

(d) α -globin expression in the isolated populations

We expected the early erythroid progenitors to be α -globin low expressing cells whereas the mature population to have high α -globin expression. Expression of the α -globin genes was assessed by Real-Time qPCR in both populations. As positive control we used FACS-sorted early progenitors and terminally differentiated cells FISH-tested for α -globin expression (Figure 39). The levels of α -globin expression in the early progenitors were comparably low to that of the control population, indicating that α -globin is transcribed at basal levels in these cells. In the differentiated cells, α -globin expression is considerably higher than in the earlier cell population, as expected.

Those two populations clearly fit the criteria for early progenitors and differentiated red cells. The HS profiles in these cell populations can now be addressed.

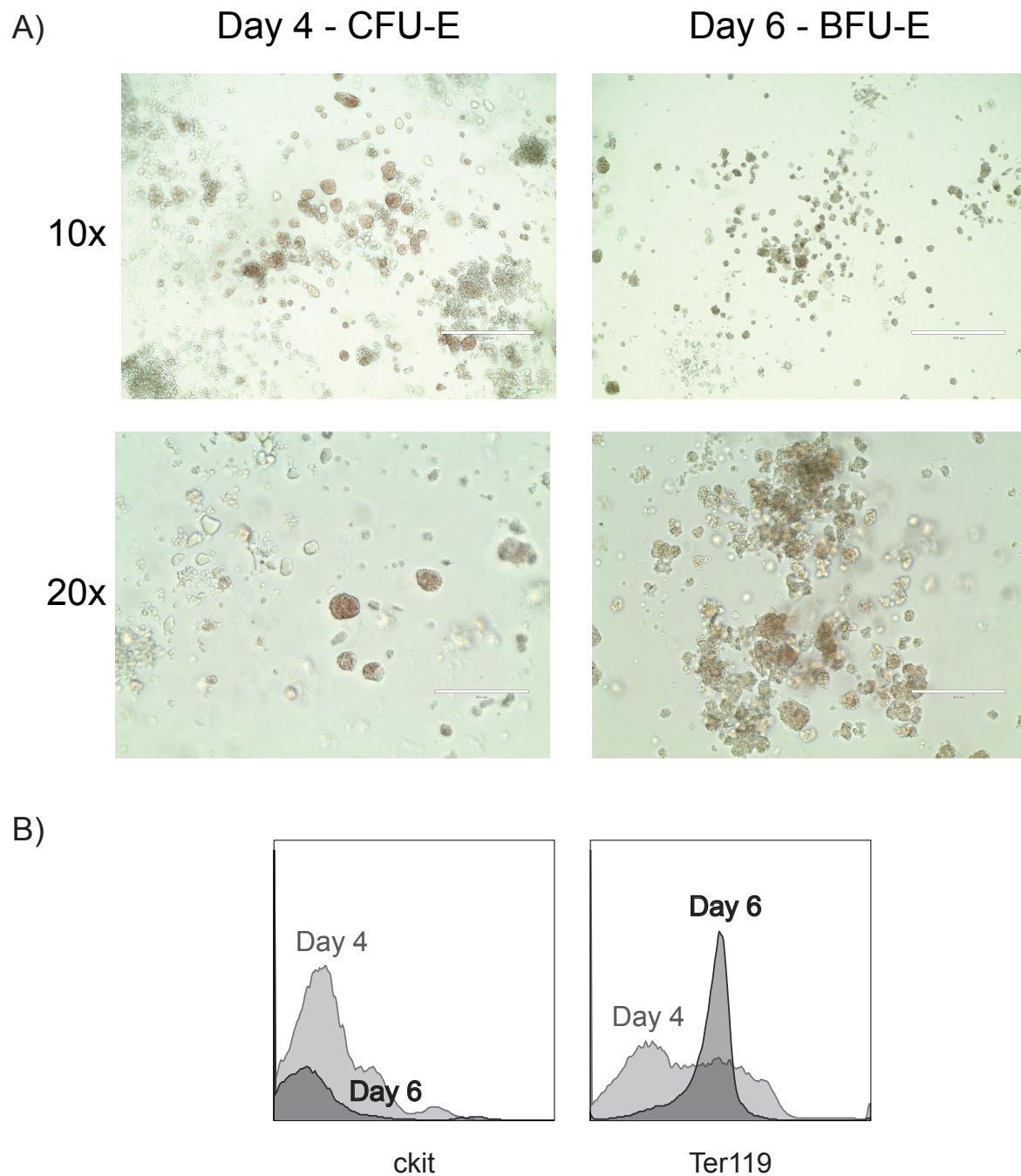


Figure 38. Developmental potential of early progenitor population (ckit+). A) Progenitor assay shows the formation of haemoglobinised CFU-E (day 4) and BFU-E (day 6) colonies derived from the early progenitor population (10X, 20X objectives 200mm scale bar). Representative photos three independent experiments. B) Histograms from the FACS analysis of cells from BFU-E and CFU-E, showing the change in the cell surface markers ckit and Ter119+ from day 4 to day 6.

α -globin expression relative to 18S RNA

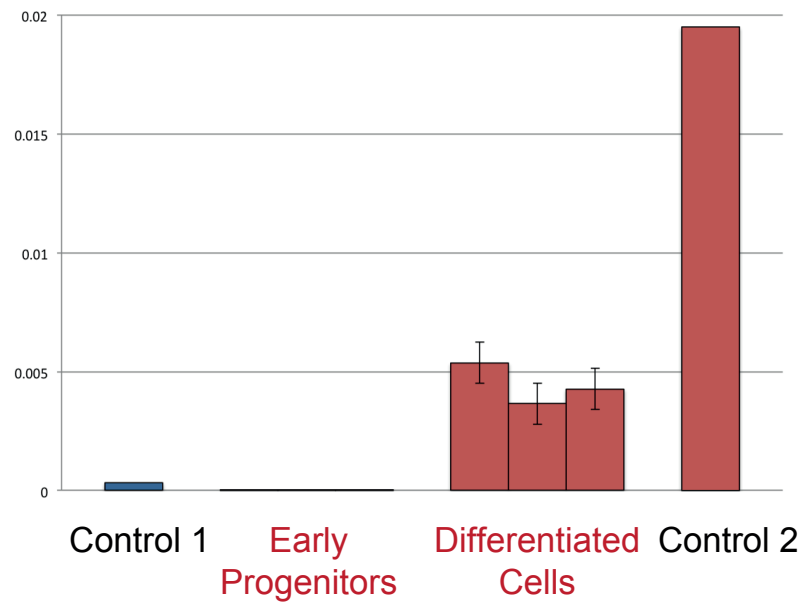


Figure 39. α -globin expression in the isolated early progenitors and differentiated populations as compared to controls by Real-Time PCR. Controls are erythroid progenitors (Control 1) and differentiated (Control 2) cells FACS-sorted and FISH tested for α -globin expression. Error bars show standard deviation from three independent experiments.

3.03 Chromatin features of the α -globin cluster differ between the two stages of differentiation

(a) The HSs in the α -globin cluster are present in the early progenitors

The open chromatin landscape was assessed using ATAC-Seq in early progenitors and differentiated cells, as well as in E14 ES cells as a non-erythroid control.

The most striking observation was that despite the promoter being transcribed at basal levels, all the α -globin regulatory regions are hypersensitive in the early progenitors. However, even if all the HS sites are formed, some differences in HS peaks are visible between the two populations. We quantified these differences by differential analysis (Materials and Methods). The results show reproducible quantitative changes (triplicates) that are statistically significant not only at the α -globin locus, but also at a genome-wide level (Figure 40).

R1, one of the two major regulatory elements, has the strongest enrichment in open-chromatin (highest peak), which does not change between the two stages of differentiation. We find that R2 and R3 hypersensitivity is significantly reduced in early progenitors, compared to differentiated cells as well as the region between the α -globin promoters and the θ -globin genes. R4 and Rm HS are not significantly different.

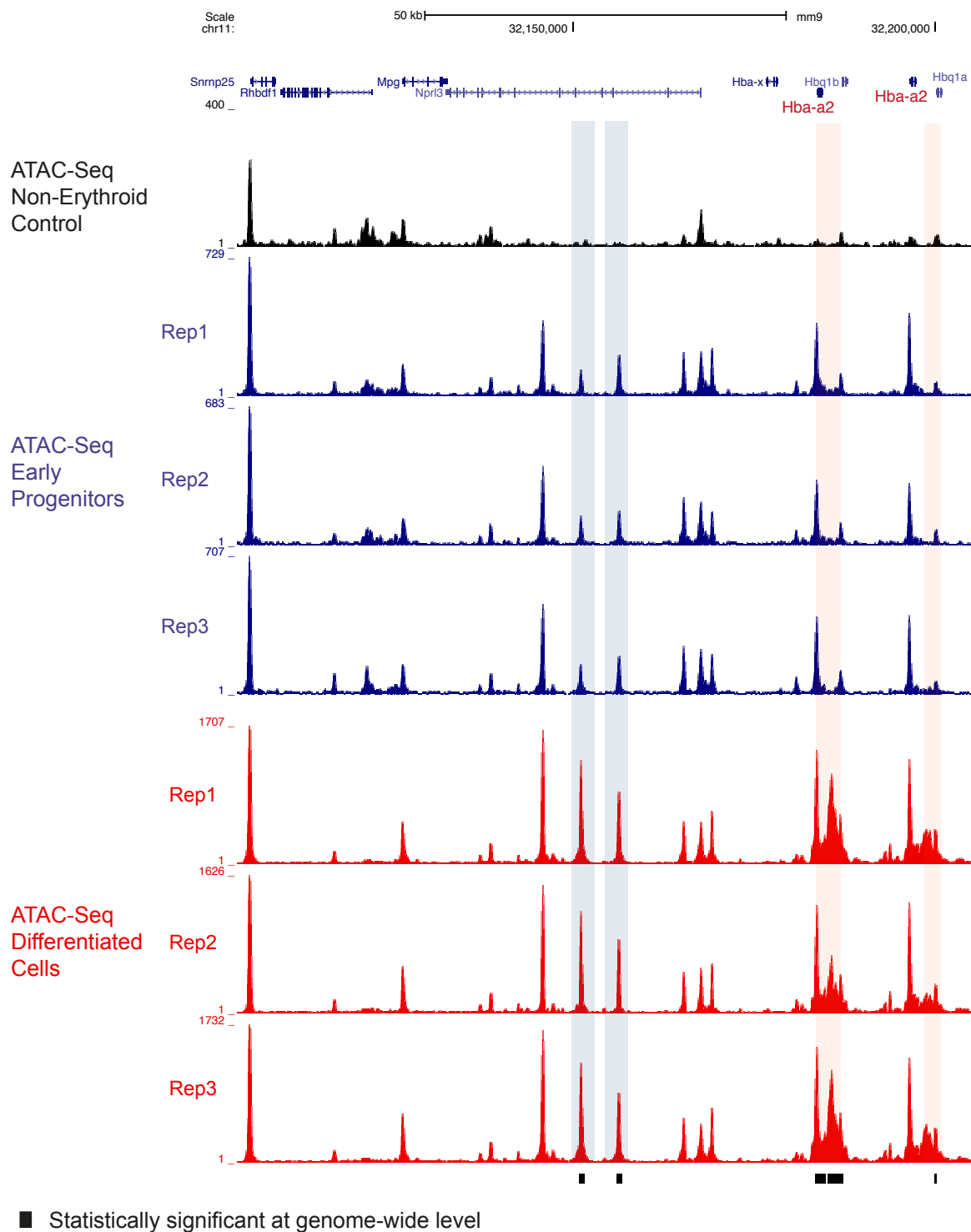


Figure 40. Open chromatin landscape at the α -globin locus in early progenitors and differentiated cells. Peaks represent areas of open-chromatin enrichment and correspond to regulatory elements previously described. The differences statistically significant obtained from quantitative genome-wide analysis are marked by black squares and highlighted in blue at enhancers) and red at promoters. Track of open chromatin in non-erythroid control (E14 ES cells invitro differentiated to day 3) shown in black. RefSeq genes shown on top.

If we assume that hypersensitivity correlates with the level of protein binding, these changes could reflect a sequential binding of proteins at the functional regulatory elements (R1, R4, Rm) followed by R2 and R3, as differentiation progresses.

(b) Chromatin marks at the α -globin locus reflect the two stages in differentiation

Histone modifications influence chromatin accessibility and protein binding. Different chromatin marks are associated with different genomic elements and are used for annotation of general *cis*-regulatory elements (H3K4me2), enhancers or promoters (H3K4me1 vs H3K4me3), and transcriptional activation or repression (H3K27Ac vs H3K27me3)^{101,119,120}. In order to analyse if any of the changes observed at the level of open-chromatin in the α -globin cluster are reflected by chromatin marks we analysed ChIP-Seq for histone marks at the two stages of erythroid differentiation and assessed enrichment (Figure 41).

We used publicly available datasets⁹² for the H3K4me1, H3K4me2, H3K4me3 and H3K27Ac marks and compared the profiles between progenitor and differentiated cell populations. The cell populations in Lara-Astiaso et al., were purified following the same immunophenotypic criteria applied in our cell isolation protocols and the use of the same antibodies and input material in the cell populations studied facilitated the comparison.

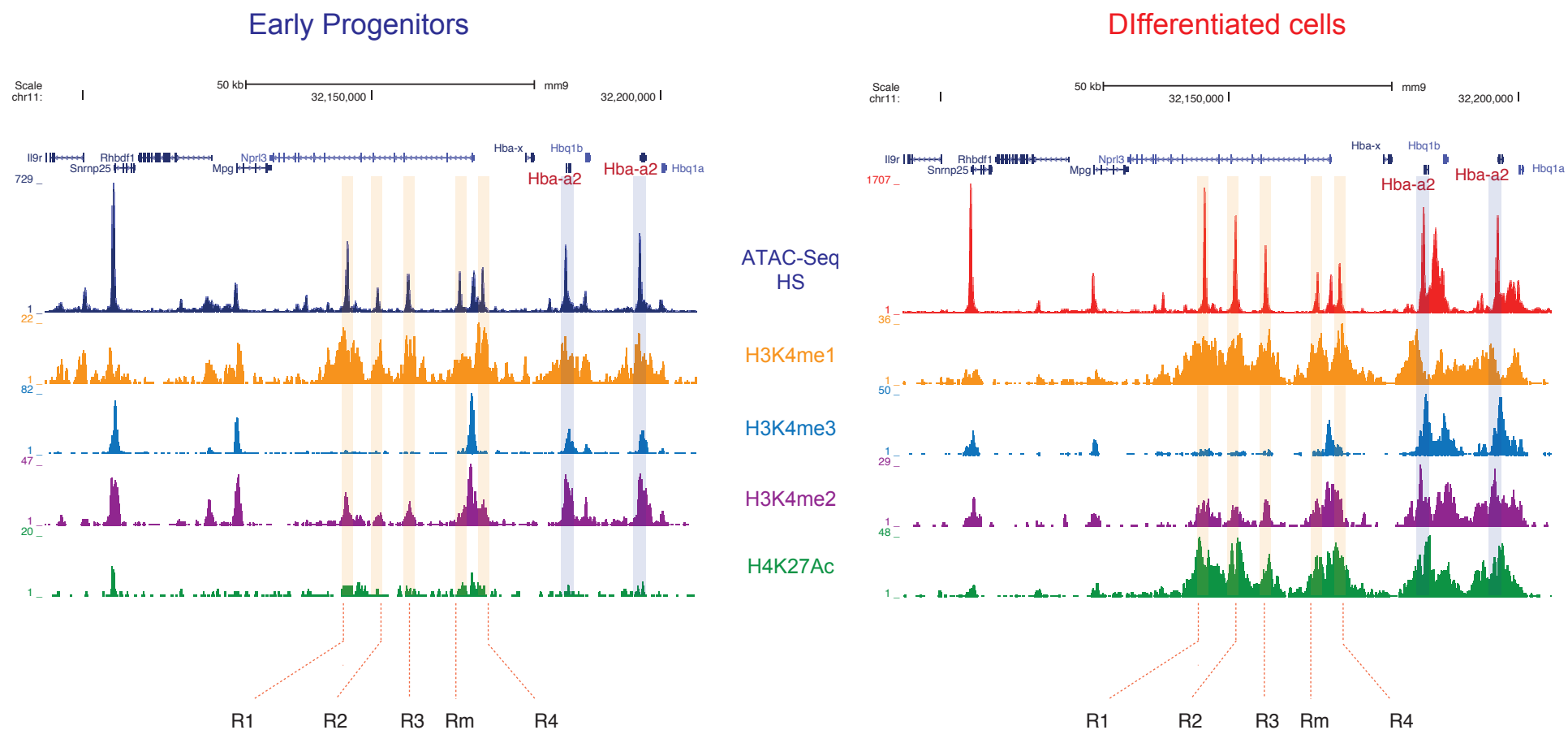


Figure 41. HS regions and chromatin marks in early progenitors and differentiated cells. UCSC tracks showing the annotated α -globin cluster in early progenitors and differentiated cells. The five enhancers (orange) and promoters (blue) correspond to HS regions as illustrated by ATAC-Seq. Publicly available ChIP-Seq for chromatin marks (Lara-Astiaso et al., 2014) illustrated for each population.

The pattern of H3K4me1 is present at the enhancers in both populations, but shows a subtle increase in signal at R2 and R3 in differentiated cells. H3K4me3 shows a different pattern at the α -globin promoters in differentiated cells, correlated with expected increase in gene activity. H3K4me2, marking both enhancers and promoters, has a similar pattern in both populations.

However, a striking difference is observed with the H3K27Ac pattern which shows low levels of enrichment in the early progenitors and much clearer, higher enrichment in the differentiated cells over the α -globin locus. This is in line with previous and recent studies and is associated with transcriptional activity at the locus in differentiated cells^{52,53,121}.

To conclude, the quantitative differences observed at the HS as the cells transition from early to differentiated cells are reflected by acetylation levels and some of the other chromatin marks: H3K4me1 is slightly increased when R2 HS is increased in differentiated cells; H3K4me3 enrichment at the α -globin promoter is higher, and has a pattern similar to the hypersensitivity as the gene becomes active.

Increased H3K4me1 levels at active enhancers, and increased H3K4me3 levels at active promoters are interesting observations which have not been reported. It is worth noting that the experiments performed in this study⁹² were limited by cell numbers and triplicates are not available. To allow for a conclusion regarding a significant difference in H3K4me1 or H3K4me3 levels, more biological replicates that allow for quantitative analysis would be required.

(c) Analysis of the chromatin interactions by Next Generation (NG) Capture-C

We next analysed the chromatin interactions in the two populations to identify potential conformational changes in the cluster that could reflect the changes in hypersensitivity and activity. We performed low number NG Capture C (Davies et al., 2015 in preparation) ¹²² in early progenitors and differentiated cells using the major regulatory elements (R1,R2) as view points (Figure 42).

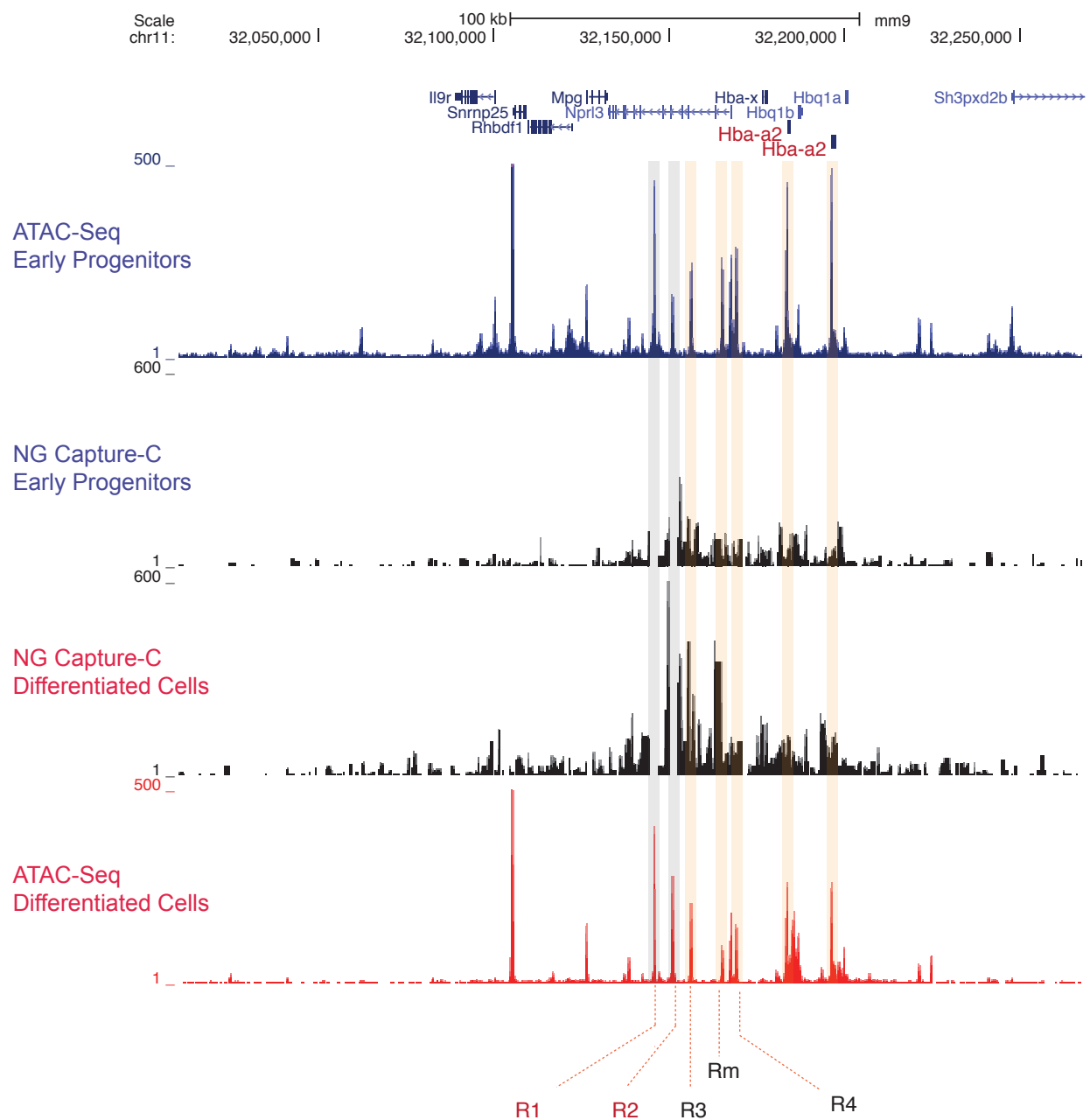


Figure 42. Comparison of the interaction profiles from the the major enhancers R1 and R2 in early progenitors and differentiated cells. NG Capture-C was performed using the R1 and R2 enhancers as a viewpoints (grey) in both early progenitors and differentiated populations (black) . Highlighted in orange are the interactions at the other enhancer elements (R3, Rm, R4) and the α -globin promoters. Shown in blue and red are ATAC-Seq tracks in the corresponding populations.

The compartment of interactions at the α -globin cluster is visible in both populations. However, the compartment is better defined and the frequency of interactions is higher in the differentiated population. New interactions with R3, R4 and the promoters can be detected as compared to the early progenitor populations. As the gene becomes transcriptionally active, in line with the increase in HS levels we observe an increase in the frequency of interactions in the same compartment.

The quality of the data is suboptimal, which could be due to the low cell numbers used. In order to draw a conclusion the NG Capture would need to be further optimised in these cells and tested from different view points, such as the promoters.

(d) Digital footprinting in the early progenitors and differentiated cells using ATAC-Seq

Direct binding of TFs at the identified hypersensitive sites was analysed by applying the in vivo footprinting method (Materials and Methods) using the same analysis principles as for DNase-Seq on ATAC-Seq data (Figure 43).

As described in Chapter 1- Digital footprints the ATAC-Seq footprints do not resolve binding sites to the same level observed in DNase-Seq footprints. Consequently, it is not possible to assess differences visually with confidence, between the early progenitors and differentiated cells. We show footprints at the enhancers R2 and R3, and the promoter where HS changes were statistically significant between the two differentiation stages.

We used the pyDNase (Chapter1) to computationally identify footprints. The results in this dataset confirm that the signal to noise ratio is too low in the ATAC-Seq data to allow for identification of footprints.

It appears that in order to compare the high-resolution protein binding profiles, DNase-Seq would need to be generated from the early progenitors, with the challenge of obtaining the required cell number (at least 10^7 cells) from such rare primary erythroid population.

The model locus perspective on TF binding during differentiation in combination with other genomic methods, offers the advantage of analysing elements which have been functionally characterised. The genome-wide perspective can reveal a signature for the differentiation process and highlight additional loci for future study in which to test principles observed at the model locus.

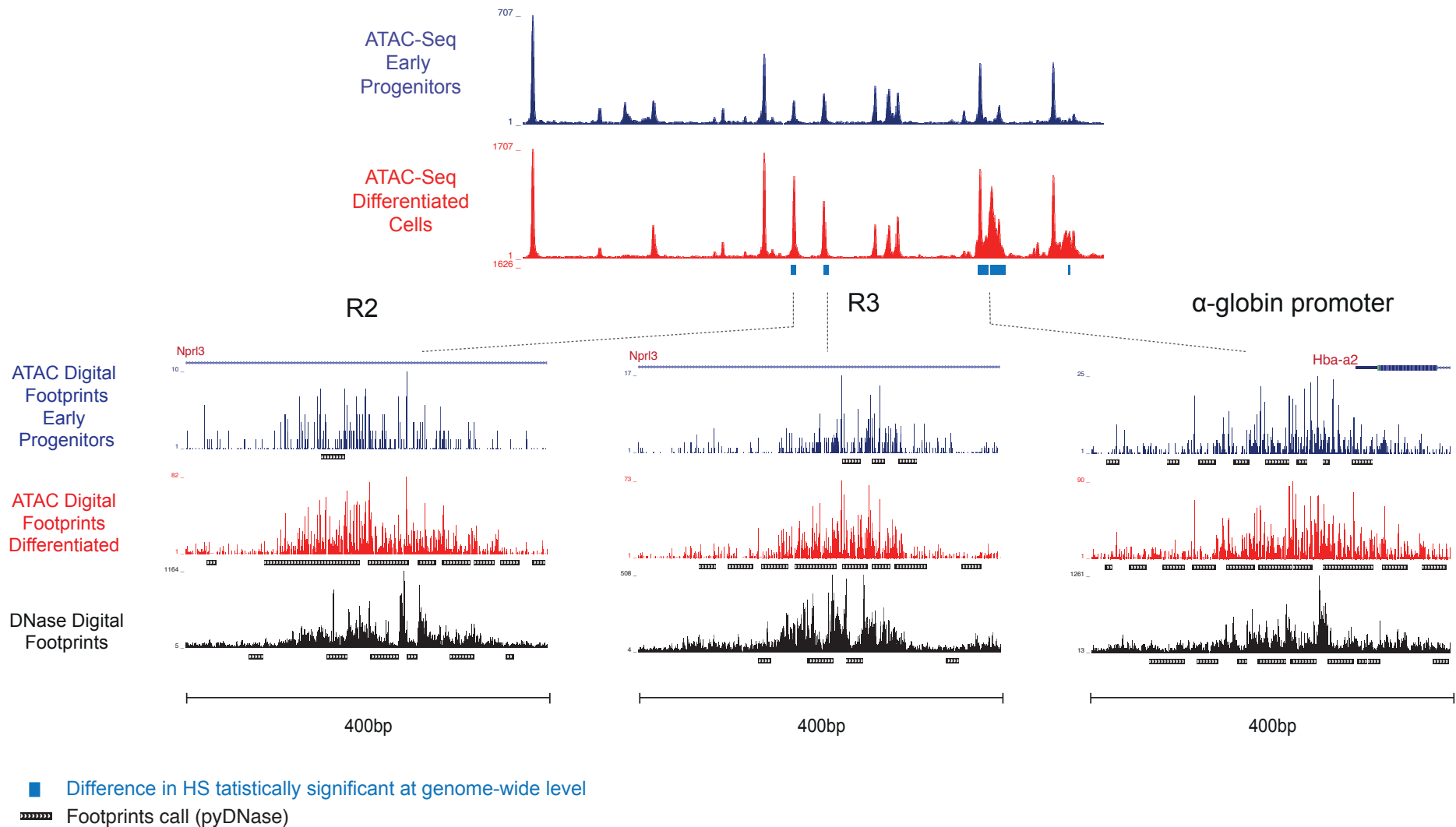


Figure 43. Digital footprints in early progenitors and differentiated cells from ATAC-Seq as compared to DNase-Seq in differentiated cells. UCSC tracks showing HSs at the α -globin cluster. ATAC digital footprints in early progenitors (blue) and differentiated cells (red) shown at the statistically significant regions- two enhancers (R2, R3) and the promoter of α -globin. DNase digital footprints in differentiated cells (black) shown for comparison. pyDNase footprint call shown below each digital footprint set.

3.04 Genome-wide analysis of HS of at the two stages of differentiation

(a) Peaks are differentially enriched at distal elements in both stages of erythropoiesis

In order to assess the global differences in hypersensitivity between early progenitors and differentiated erythroid cells, we performed genome-wide differential analysis (R package, materials and methods). The analysis of triplicates with methods designed for genome-wide analysis allowed us to quantify the differences in open chromatin and highlight the statistically significant changes.

Peaks corresponding to the HS regions were identified and filtered resulting in sets of 25043 open-chromatin peaks in early progenitors and 22329 peaks in differentiated cells. Differential analysis of the enrichment at these sites reveals that 2656 (12.1%) peaks are significantly enriched in early progenitors, and 2175 (9.9%) in differentiated cells. 78% of sites do not change (above the significance threshold) between the two cell types (Figure 44). It is interesting to note that the majority of the open-chromatin landscape does not change; 10-12% of the HS regulatory regions are sufficient to drive the differences between the two stages of erythropoiesis.

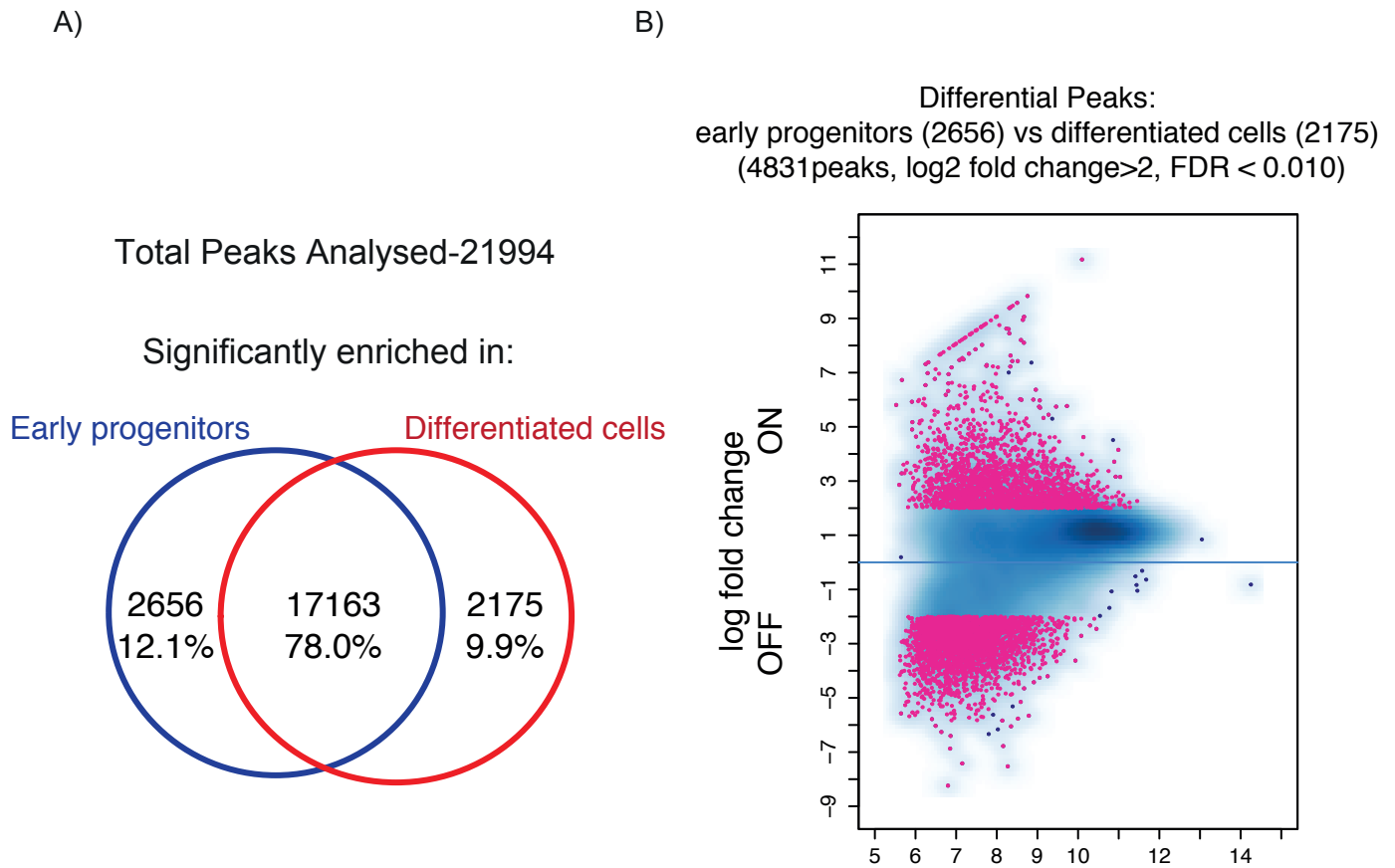


Figure 44. Overlap and differential peaks in early progenitors and differentiated cells (AT-AC-Seq). A) Venn diagram illustrating the proportion of peaks, out of total, that are differentially enriched (log2 fold change > 2 and FDR < 0.01) in progenitor and differentiated cells. B) Fold changes in read density (log2) between two samples (y-axis) are plotted against the average of read density (x-axis) at the same peaks. Differential peaks in the differentiated population will have a positive log fold change, while differential peaks in the progenitor population will have a negative fold change. Differential analysis performed with three independent biological replicates.

(b) Chromatin marks at the cell-specific regions

To characterise the differentially enriched peaks, we annotated these regions using chromatin marks. To allow for a comparison in using the same antibodies and similar cell numbers, we used the publicly available ChIP-Seq data for the annotation of early erythroid progenitors and differentiated cells (as mentioned above in Section 3b) ⁹².

We generated heatmap plots of the signal of the chromatin marks at the total and differential peaks in the two populations (Figure 45). At the total set of peaks, the signal of chromatin marks annotate more promoters than enhancers. Acetylation and di-methylation levels can be observed at both enhancers and promoters.

In the majority of “differentiated-only” peaks (almost 80%) are enhancers. Thus, at a genome-wide level, the cell type specific regulatory regions, enriched in open-chromatin correspond mainly to distal elements. In the “progenitor-only” set, based on H3K4me1 and H3K4me3 marks, the distal regions contain more enhancers than promoters though the distribution is less striking than in the “differentiated-only” peaks.

We compared the other marks, H3K27Ac and H3K4me2 at the differential peaks. The acetylation levels appear stronger at the “differentiated-only” regions, whereas di-methylation appears stronger at “progenitors-only” set.

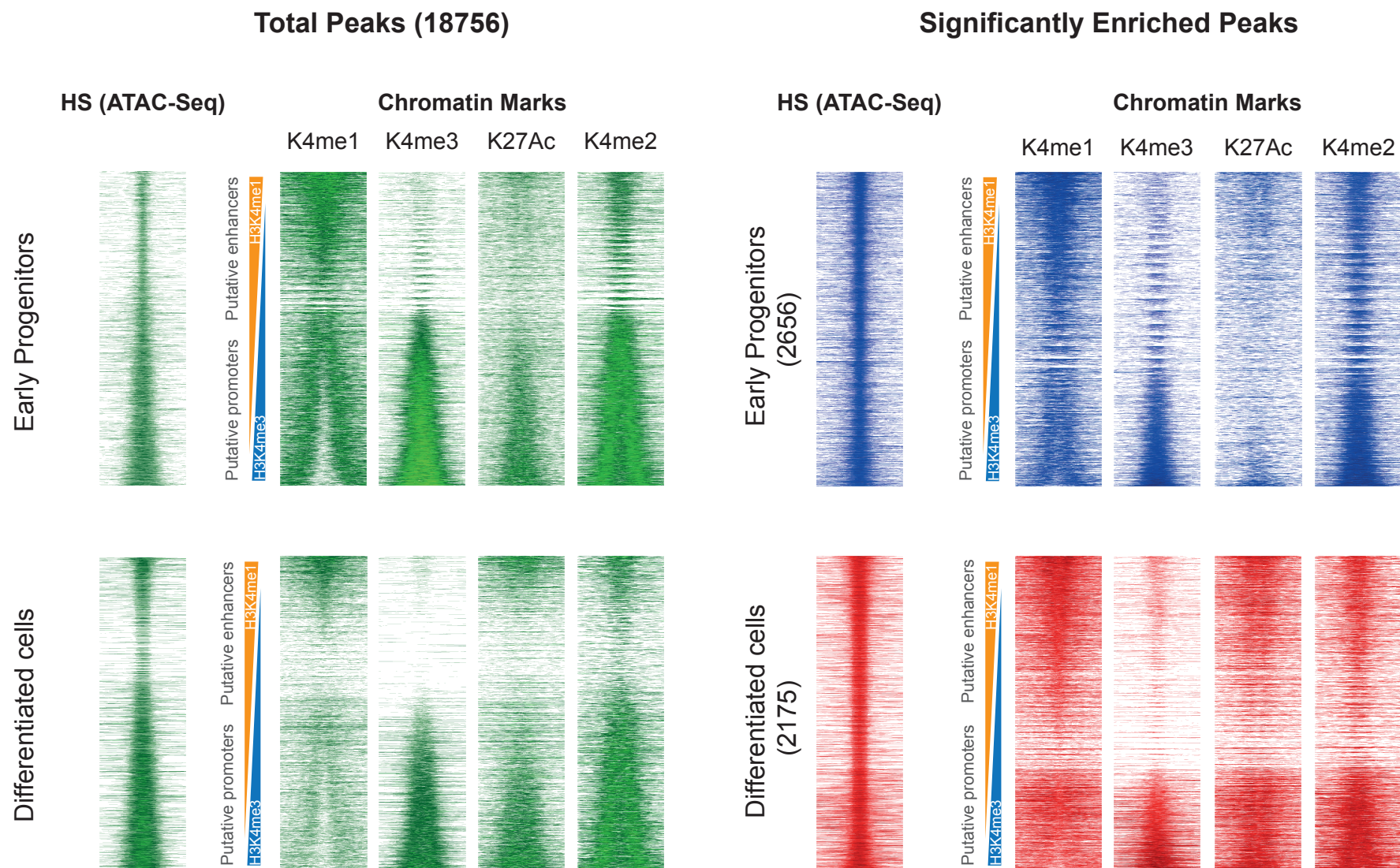


Figure 45. Heatmap representation of ATAC-Seq and ChIP-Seq in early progenitors and differentiated cells. Signal 4kb around the total peak call regions, or significantly differential peaks, sorted by the difference in H3K4me1 and H3K4me3 enrichment. In the side panes the annotated "putative enhancers" are marked orange, and "putative promoters" marked blue. Enhancers and promoters are identified as hypersensitive regions with more than 10 reads per million (RPM) and a H3K4me1 to H3K4me3 difference in normalised reads higher than 1 ($H3K4me1 > H3K4me3$) or lower than 1 ($H3K4me3 < H3K4me1$) respectively.

(c) Examples of statistically significant regions in the two populations

A few examples were selected to illustrate the differential regions in the two populations.

In early progenitor cells (Figure 46), loci such as *Kit* (gene expressing the early progenitor cell surface marker cKit) and *Gata2* show enriched hypersensitivity at enhancers and promoters marked by corresponding chromatin marks. These loci lose most of their hypersensitivity in differentiated cells. *Notch2* (involved in control of cell growth and differentiation)¹²³ maintains an open promoter in both populations, but tissue specific enhancers are visible in the early progenitors. In contrast *Bcl2*, (prevents apoptosis in early progenitors)¹²⁴, loses its HS regions by late differentiation.

In differentiated cells (Figure 47) we find significant enrichment at genes encoding membrane proteins such as anion exchanger *Slc4a1*. A GATA1 target, *Gstp2*⁶² becomes hypersensitive in differentiated cells, together with an intronic enhancer.

We also show the β -globin and the mitoferrin1 (*Slc25a37*) loci, where apart from the increase in hypersensitivity in differentiation cells, we detect a similar chromatin mark signature to the one observed in the activation of α -globin expression: marked increased in K27Ac, increased H3K4me1 at enhancers and increased H3K4me3 at promoters.

To summarise, by quantifying and identifying the cell-type specific hypersensitive regions we isolate the specific regulatory elements that most likely drive the differences between the two stages. Majority of these cell-specific HSs are found in different genomic locations between the two cell types, are enriched in chromatin marks and correspond predominantly to enhancer elements.

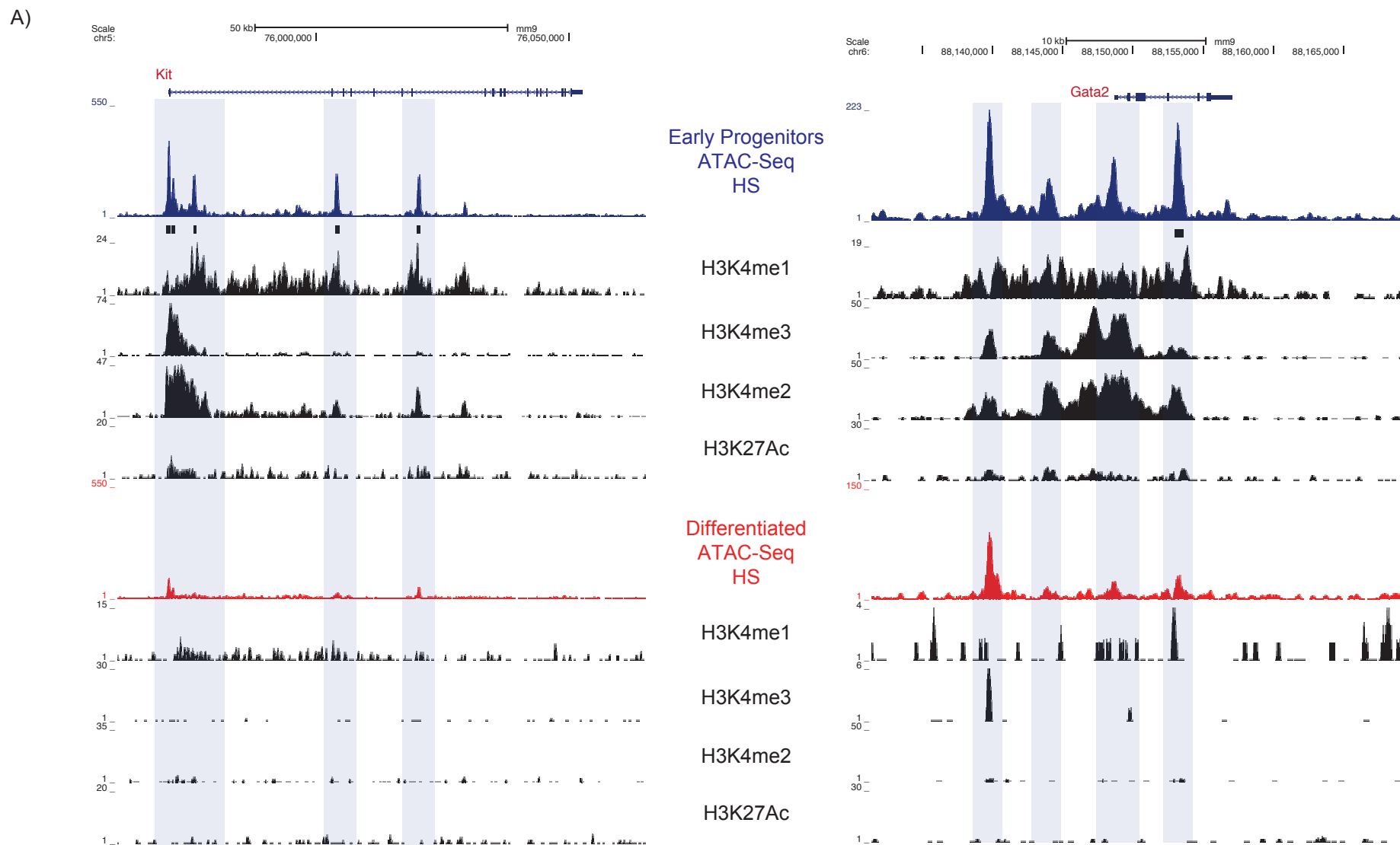
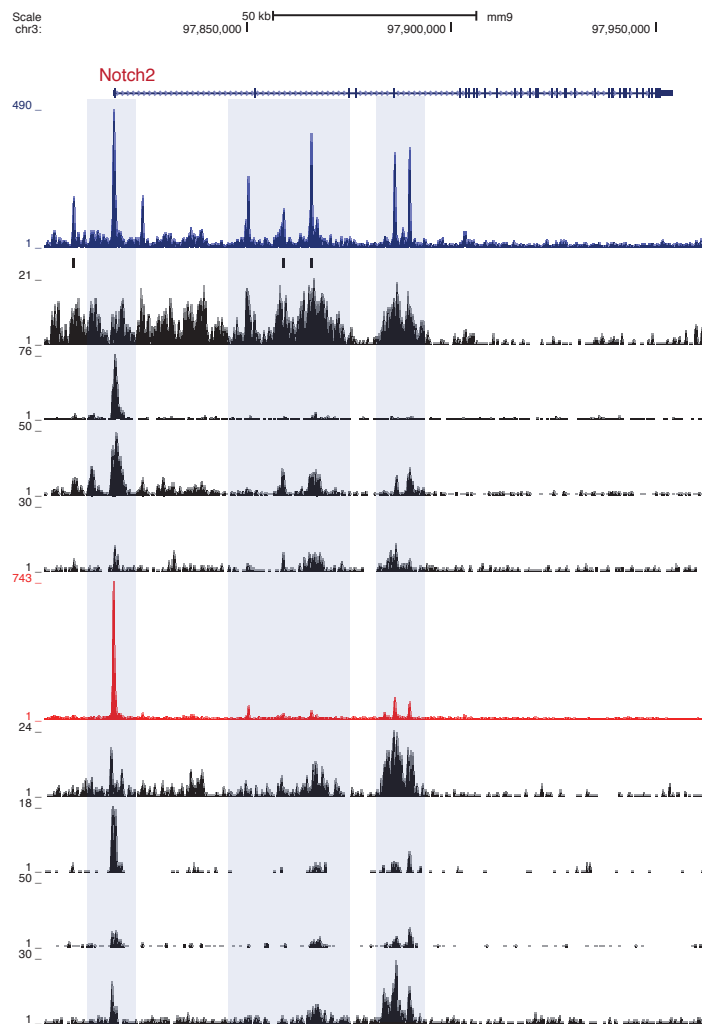


Figure 46. Examples of significantly enriched HS regions in early progenitors. UCSC tracks showing HS regions from ATAC-Seq and chromatin marks from ChIP-Seq (Lara-Astiaso et al., 2014) at selected genomic locations in early progenitors (blue) and differentiated cells (red). A) *Kit* and *Gata2*; B) *Notch2* and *Bcl2*. Highlighted regions correspond to elements (enhancers or promoters) enriched in early progenitor cells. Black boxes mark regions with statistically significant enrichment. RefSeq genes are shown on top.



Early Progenitors
ATAC-Seq
HS

H3K4me1

H3K4me3

H3K4me2

H3K27Ac

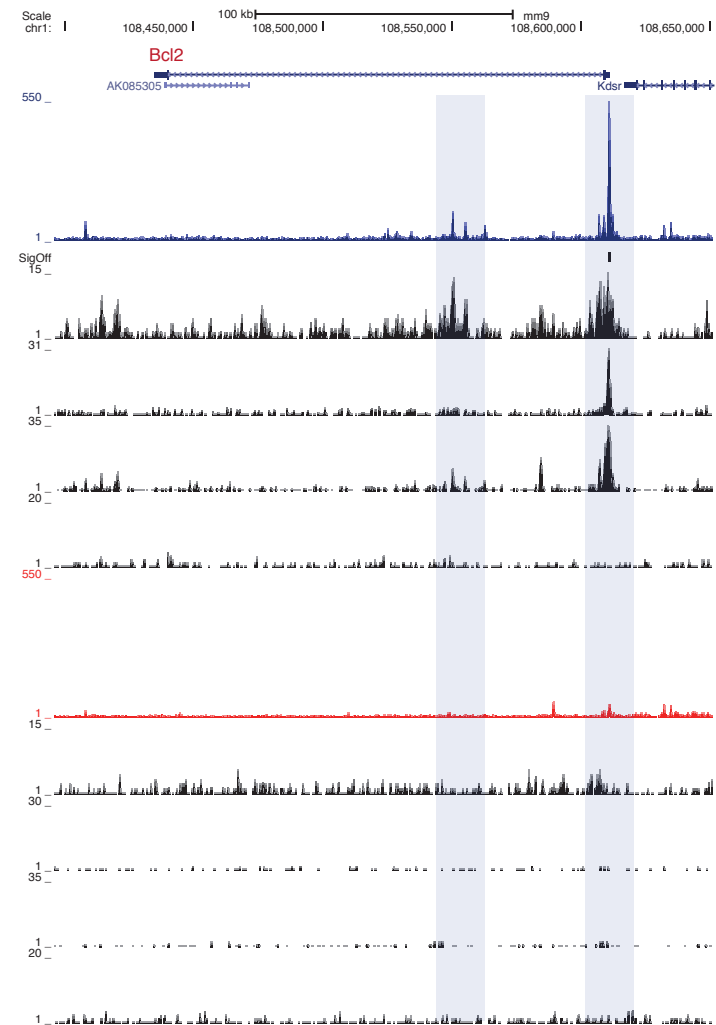
Differentiated
ATAC-Seq
HS

H3K4me1

H3K4me3

H3K4me2

H3K27Ac



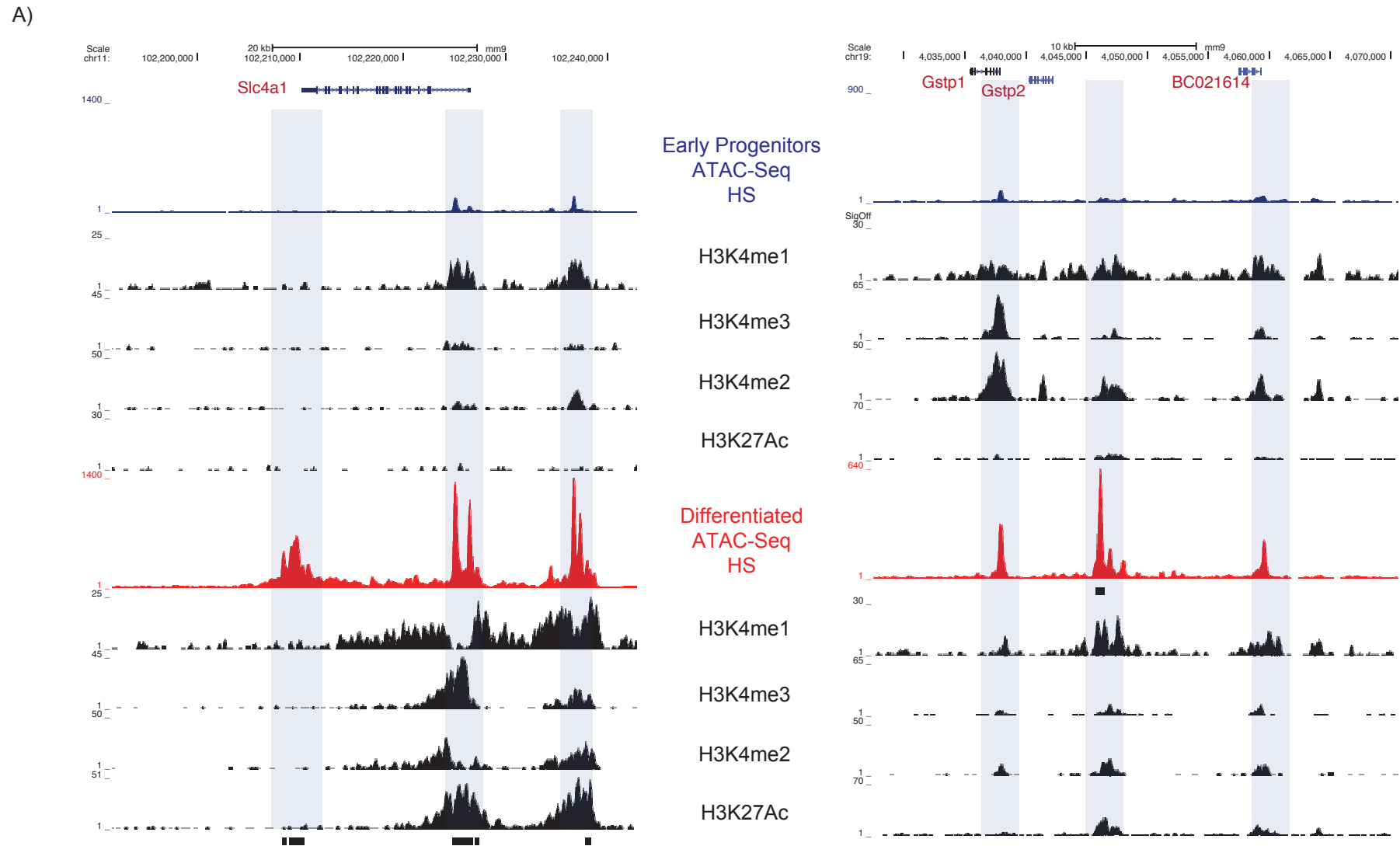
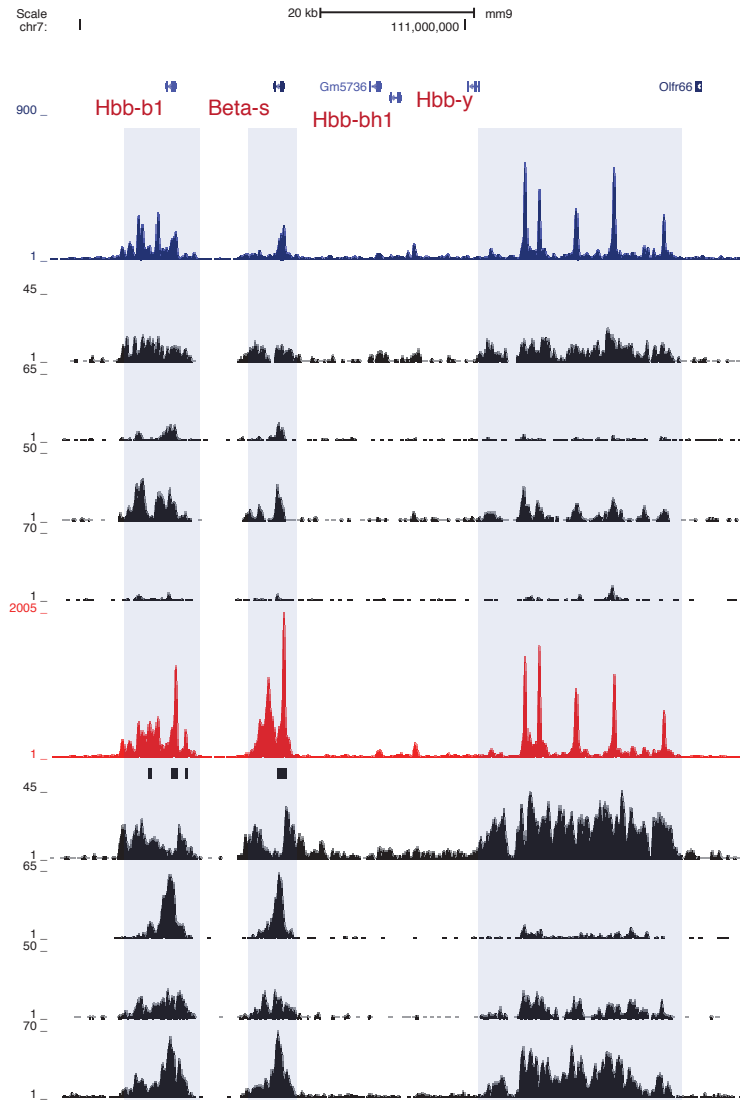


Figure 47. Examples of significantly enriched HS regions in differentiated cells. UCSC tracks showing HS regions from ATAC-Seq and chromatin marks from ChIP-Seq (Lara-Astiaso et al., 2014) at selected genomic locations in early progenitors (blue) and differentiated cells (red). A) *Slc4a1* and *Gstp2*; B) *Hbb* and *Slc25a37*. Highlighted regions correspond to elements (enhancers or promoters) enriched in differentiated cells. Black boxes mark regions with statistically significant enrichment. RefSeq genes are shown on top.

B)



Early Progenitors
ATAC-Seq
HS

H3K4me1

H3K4me3

H3K4me2

H3K27Ac

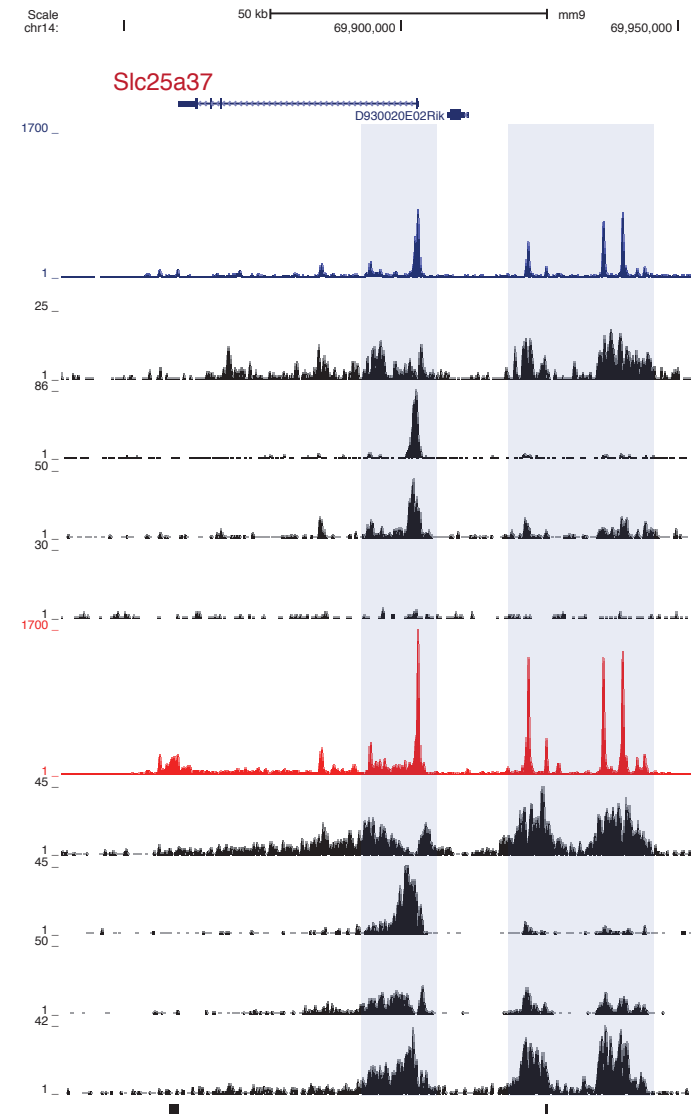
Differentiated
ATAC-Seq
HS

H3K4me1

H3K4me3

H3K4me2

H3K27Ac



3.05 Conclusion

The differences in hypersensitivity reflect changes in activity of the α -globin locus, though the contrast is not as striking as we expected. Even though the promoters are transcribed only at basal levels in the progenitor population, all the HSs corresponding to the enhancers are already formed.

However, we find interesting quantitative differences in the level of open chromatin enrichment. The level of α -globin activity is reflected by the degree of HS of some of the enhancers (R2, R3). The differences in HS are also reflected by chromatin marks and interactions. At the locus, the distinctive histone modifications between the early progenitors and differentiated cells, are H3K27Ac overall, but also levels of H3K4me3 at promoters, and H3K4me1 at enhancers.

The direct TF binding at these HS regions could not be assessed, as ATAC-Seq signal does not resolve digital footprints at the same resolution as DNase-Seq

We have in addition to the model locus, also analysed the two stages genome-wide to identify the differences in HSs associated with differentiation. The differences between the early progenitor and the differentiated population are found at regulatory elements, mainly enhancers. Thus, cell specific differences in terms of bound regions lie within distal enhancers.

In order to identify differences in TF binding at these regulatory sequences, it would be interesting to analyse the global changes in footprints at these sites. The feasibility of this aspect will be addressed in Results- Chapter 4.

Chapter IV. A novel approach for the analysis of footprints

4.01 Introduction

High resolution analysis of hypersensitivity data was the main aim of my study. It offers a powerful tool to resolve sequences that are bound by known or unknown proteins and which could potentially be dictating the nature of the cells at the different stages analysed.

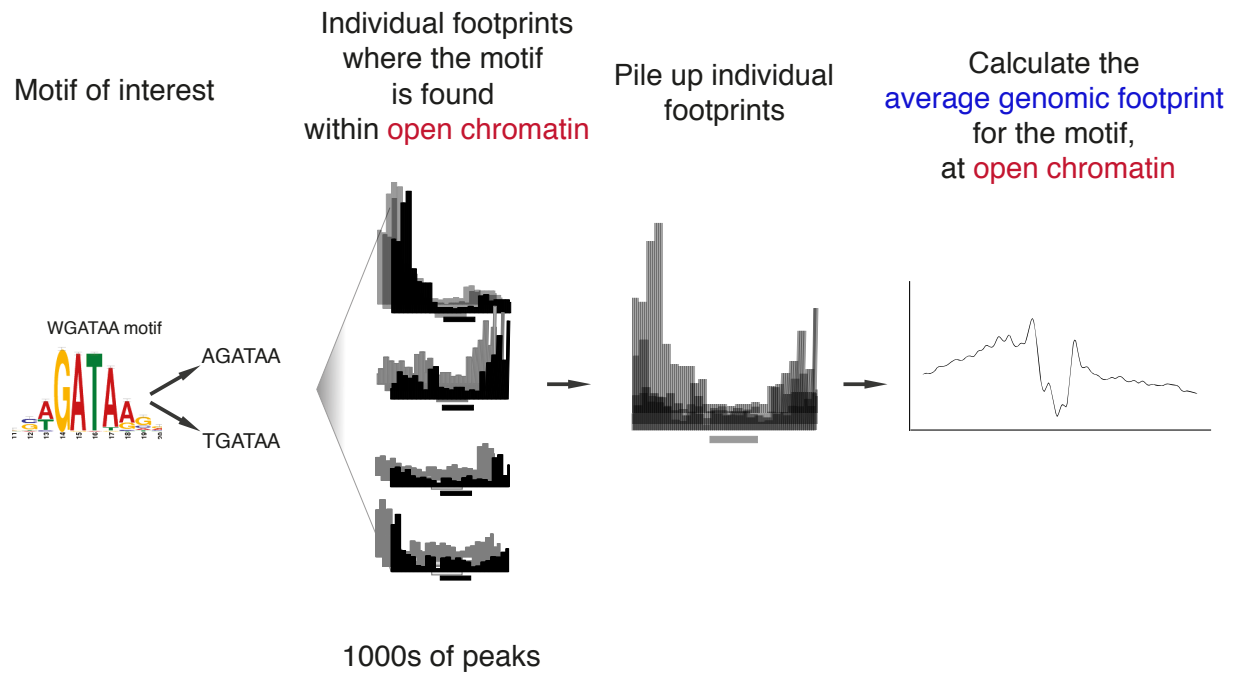
We showed that using DNase digital footprints, we can compare the direct binding of TFs at the α -globin locus in primitive and definitive cells. With the scope of unravelling patterns of direct binding and infer sequence usage in different cell types, in this chapter, we took a different approach to analysing DNase-Seq digital footprinting data.

Instead of analysing single digital footprints, the new approach takes into account the genome-wide HS regions, averages the digital footprints for a specific sequence, and scores them. Starting with DNase-Seq data, we generate all the average footprints from primitive and definitive erythroid cells and show a comparison of the most used sequences in the two cell populations. We test the “average footprinting” approach on ATAC-Seq data. Finally, we compare and contrast DNase and ATAC average footprinting data.

4.02 Computing genome-wide footprints - ‘Average’ footprints from DNase-Seq data

The new approach for generating “average footprints” involves generating an average profile for any 6 bp motif occurring in a hypersensitivity dataset of interest. (Figure 48 A): all the HSs are scanned for the motif of interest; when the motif is found, the digital footprint is extracted and stored. Finally, all the stored footprints for a motif are piled up into an average profile. This generates an “average footprint” for the motif queried, in the cell type of interest. The average footprint can be scored, based on the difference between the frequency of cuts at the motif location (black lines) and motif margins (red lines) (Figure 48B). The average footprint score represents an average of the depth of the footprint from all the occurrences in the data. The ratio of occupied sites to unoccupied sites which contain the motif is reflected in the footprint depth- the more occupied, the stronger the depth, and higher the score.

A)



B)

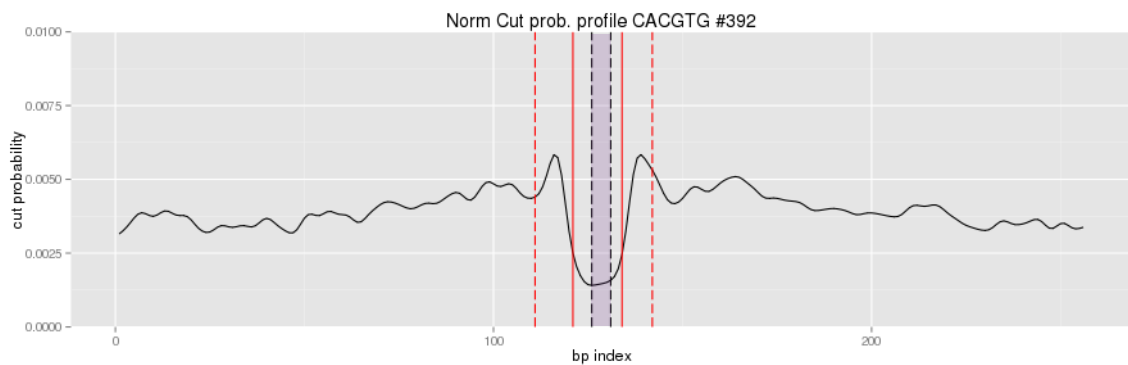


Figure 48. Generation of genome-wide footprints. A) Diagram illustrating the approach to generate the footprint average profile of a specific motif. B) Example of an average profile, in this case the E-box motif **CACGTG**. The profile was generated from DNase-Seq data from definitive erythroid cells. The highlighted dotted black line shows the location of the motif within the footprint. The red dotted lines represent the areas determined by the algorithm to be the margins of the footprint ("shoulders").

For an unbiased exhaustive analysis, we expanded the approach and analysed all possible 6 bp motifs occurring in the genome (Figure 49). For six-letter combinations of a four-letter genome, there is a finite number of motifs (4096) that can be generated and tested. The average footprints for all these motifs can be calculated and ranked depending on their score, resulting in a comprehensive list of sequences that are bound, and thus form footprints, in the cell type of interest.

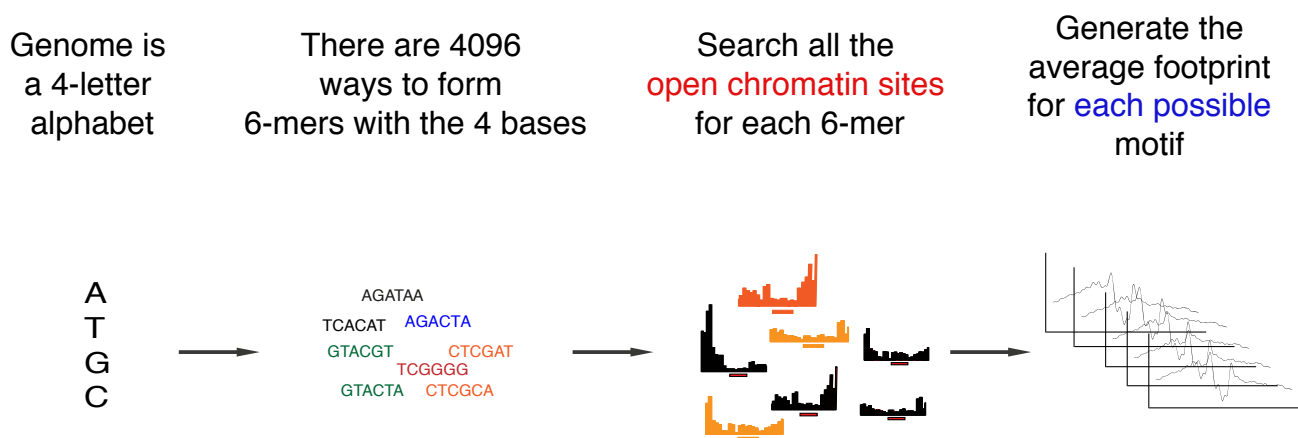


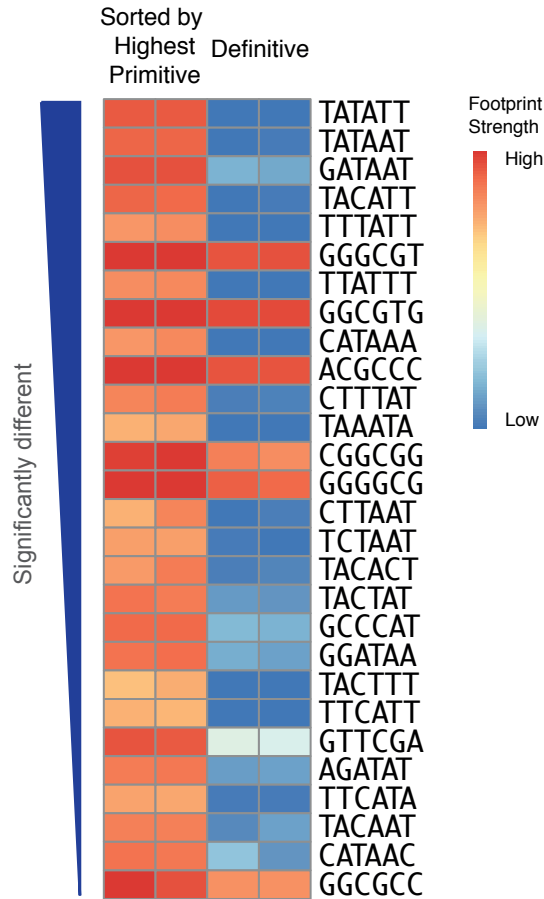
Figure 49. Identifying average footprints for any 6bp motif. Diagram illustrating the approach to generate average footprints profile of any possible 6bp. All the 6-letter combinations using the 4bp are calculated. Next each of these motifs is used to generate and average footprint in the cell type of interest. The profile of the footprint as well as the score are calculated.

(a) Average footprints in primitive and definitive cells

We generated the average footprints for all possible 6bp motifs in primitive and definitive erythroid DNase-Seq data. Each average footprint was extracted, scored based on their occurrence at the total HS regions and ranked. Considering the motifs are assessed at the same genomic regions, and thus will have the same number of occurrences for both tissues, the rank of each motif within the 4096 possibilities interrogated, reflects the motifs' usage in that cell type.

The result is a list per cell type, containing all 6-bp sequences that are bound, sorted by strength of footprint to reflect usage. We analysed the motifs that have significantly different ranks in the two cell types: motifs that do not form a footprint and thus rank low, versus motifs that form footprints and rank high. We identified all the motifs that significantly change their ranks between the two cell types, suggesting they are bound in more instances in one cell type than the other (Figure 50 & 51).

A)

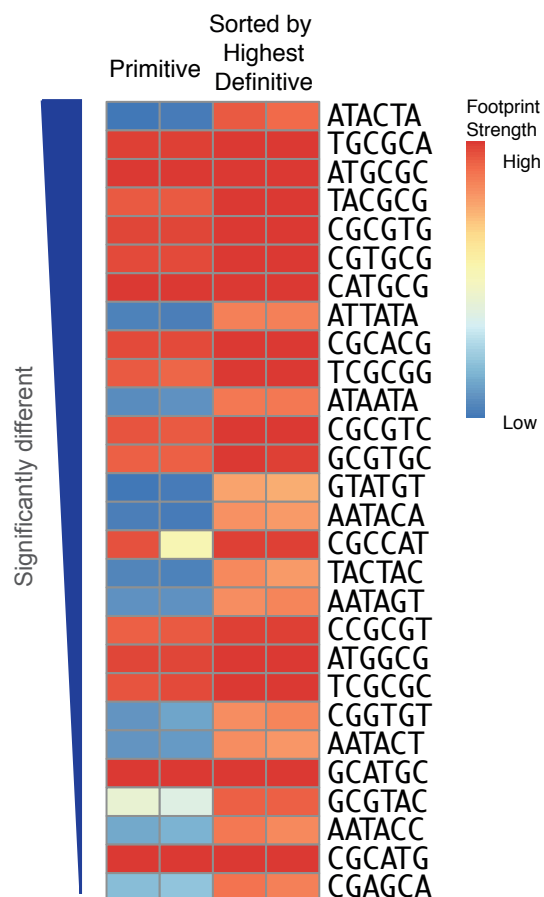


B)

Motif	Jaspar Prediction	FootprintDB Prediction
TATATT	FOXL1=0.74; NKX3-1=0.62; SRY=0.54; FOXD1=0.53; FOXC1=0.51; GATA3=0.51; SPIB=0.5;	Pit1, Prop1, Irx3, Zfp128, Glis2, Foxl1, Arid5a, NKX31, HMGA1
TATAAT	FOXL1=0.66; SRY=0.58; NKX3-1=0.56; SOX9=0.55; FOXH1=0.51; CDX2=0.51; GATA3=0.51;	Sox30, Sry, Sox14, Sox21, Sox1, Tcf7, ZN333
GATAAT	GATA3=0.63; SRY=0.57; NKX3-1=0.56; SOX9=0.54; FOXL1=0.53; CDX2=0.51; FOXH1=0.5;	Pou6f1, ZN333, Dobox5, Gata3, Gata2, Gata4, Gata6
TACATT	FOXC1=0.76; FOXL1=0.64; NKX3-1=0.62; SPIB=0.6; SOX9=0.55; DUX4=0.53; ZNF354C=0.51; SRY=0.51;	Foxk1, Foxl1, Sox17, Foxa2, Sox15, Zfp187, TEAD1
TTTATT	FOXL1=0.65; GATA3=0.62; SRY=0.6; NKX3-1=0.57; CDX2=0.56; FOXD1=0.56; FOXP2=0.54; FOXH1=0.53; BRCA1=0.52; FOXI1=0.52; SPIB=0.52; SOX9=0.52; STAT3=0.5;	HXC6, Hoxd13, Hoxb13, Hoxa13, Foxq1, Foxf1, Hoxc10
GGGCGT	FOXC1=0.66; NFIC=0.65; THAP1=0.62; KLF5=0.59; MZF1_1-4=0.59; ZNF354C=0.59; HINFP=0.56; REL=0.52; BRCA1=0.52; NRF1=0.5; FOXL1=0.5;	EGR1, KLF12, KLF7, ZIF268, KLF4, Egr3,
TTATTT	NKX3-1=0.72; FOXD1=0.65; GATA3=0.63; FOXL1=0.63; SRY=0.62; SPIB=0.59; SOX9=0.57; FOXP2=0.53; FOXH1=0.51; BRCA1=0.51; FOXI1=0.51; FLI1=0.5; CDX2=0.5;	Hoxa13, Foxa2, Foxf1, Foxq1, Foxc1, Mtf1, FOXJ2, FOXJ3, PO3F2, HMGA1
GGCGTG	BRCA1=0.64; FOXC1=0.63; NFIC=0.62; ZNF354C=0.59; KLF5=0.59; SREBF1=0.59; SREBF2=0.59; ZEB1=0.53; MZF1_5-13=0.52; THAP1=0.51;	KLF4, KLF12, Sp4, EGR1, Arnt, Ahr, ZIF268, DSNR
CATAAA	GATA3=0.62; FOXC1=0.59; FOXL1=0.57; CDX2=0.56; FOXD1=0.56; SPIB=0.55; SRY=0.54; BRCA1=0.54; BATF::JUN=0.52; STAT1=0.51; STAT3=0.51; MZF1_1-4=0.51;	HXA10, CDX1, PO6F1, Hoxc9, Hoxa9, Hoxd9, TEF, PO3F2, CDX2, Hoxc10
ACGCC	FOXC1=0.66; NFIC=0.65; THAP1=0.62; KLF5=0.59; MZF1_1-4=0.59; ZNF354C=0.59; HINFP=0.56; REL=0.52; BRCA1=0.52; NRF1=0.5; FOXL1=0.5;	EGR1, ZIF268, KLF4, KLF12, KLF7
CTTTAT	GATA3=0.72; FOXC1=0.62; FOXL1=0.6; BRCA1=0.6; RORA_1=0.55; CDX2=0.55; ZNF354C=0.55; FOXD1=0.54; SPIB=0.53; EHF=0.51; SOX9=0.51;	Glis2, CDX1, CDX2, Foxc1, Hoxa13, E2F2, Hoxd13, Sox2, TBP
TAAATA	NKX3-1=0.78; FOXL1=0.73; FOXD1=0.69; SRY=0.6; SPIB=0.56; FOXC1=0.55; BRCA1=0.55; FOXI1=0.52; FOXH1=0.52; FOXP2=0.51; GATA3=0.51; CEBPA=0.5;	Foxc1, MEF2C, PO3F2, Lhx5, FOXJ2, MEF2A, HMGA1, HNF1B, Hoxc1
CGGCGG	MZF1_1-4=0.7; NFIC=0.58; FOXC1=0.57; BRCA1=0.55; ELK1=0.55; E2F4=0.54; SPIB=0.53; THAP1=0.52; ZNF354C=0.52; E2F6=0.52; NRF1=0.5; MZF1_5-13=0.5;	EGR4, MBD2, ZIF268, SP1, EGR1, KLF6, EGR2, E2F6, TYY1
GGGGCG	MZF1_1-4=0.69; NFIC=0.66; ZNF354C=0.66; THAP1=0.62; KLF5=0.6; FOXC1=0.6; SP1=0.55; E2F1=0.55; BRCA1=0.54; MZF1_5-13=0.54; EBF1=0.52; ZEB1=0.51; REL=0.5; SREBF1=0.5;	KLF4, Sp1, KLF6, WT1, EGR1, Bcl6b, Sp4
CTTAAT	FOXC1=0.63; GATA3=0.6; NKX3-1=0.58; FOXL1=0.58; ZNF354C=0.55; RORA_1=0.54; SRY=0.53; DUX4=0.53; CEBPB=0.52; CEBPA=0.52; FOXH1=0.52; EHF=0.51; SPIB=0.51;	Nkx2, ARI3A, Pou1f1, Hoxa3, Vax2, Pou2f1
TCTAAT	GATA3=0.62; FOXL1=0.6; SPIB=0.6; NFIC=0.58; DUX4=0.54; FOXH1=0.52; ZNF354C=0.51; SRY=0.5;	Alx1, Pdx1, Zbtb12, Hxd13, Msx-1, Pou2f3, Ari3a, Pdx1, Pou2f2
TACTAT	FOXC1=0.75; ZNF354C=0.68; SPIB=0.66; NKX3-1=0.62; FOXL1=0.59; MZF1_1-4=0.58; SOX9=0.55; FOXH1=0.53; SRY=0.51; ZEB1=0.5; THAP1=0.5;	Irf2, Foxg1, Gabpa, Gbx1, Sox10, Sox11, Lhx6
TACTAT	FOXC1=0.69; SPIB=0.65; FOXL1=0.64; NKX3-1=0.62; SOX9=0.58; MZF1_1-4=0.56; NFIC=0.52; SRY=0.51;	Foxj1, Zfp187
GCCCAT	ZNF354C=0.73; NFIC=0.71; MZF1_1-4=0.65; THAP1=0.65; FOXL1=0.57; SREBF2=0.56; SPIB=0.56; SREBF1=0.56; MZF1_5-13=0.55; SP1=0.52; EBF1=0.52; CDX2=0.51; NRF1=0.51;	SRF, ZIF268, DSNR ZFP, Klf12, QGSR ZFP
GGATAA	ZNF354C=0.65; GATA3=0.63; SPIB=0.61; MZF1_1-4=0.57; NFIC=0.56; FOXL1=0.55; THAP1=0.54; SRY=0.52; SOX9=0.52; REL=0.51; FOXP2=0.5;	Gata3, Gata2, Gata6, Gata4, Gata1, NFATC2
TACTTT	FOXC1=0.79; NKX3-1=0.72; SPIB=0.67; FOXL1=0.62; DUX4=0.52; BRCA1=0.5; SOX9=0.5;	HXC6, NR1D1, Myt1, Foxj3, RORG, FOXF2, FOXA2, Nr2e1, HMGA1
TTCATT	SPIB=0.63; FOXL1=0.6; FOS=0.53; BATF::JUN=0.52; SOX9=0.52; DUX4=0.52; JUND=0.52; FOXC1=0.51; THAP1=0.51; ZNF354C=0.51; EHF=0.51; NKX3-1=0.5;	Hbp1, Bbx, Sox15, Sox17, PIT1, Irf4, ZN333
GTTCGA	BRCA1=0.58; NFIC=0.58; SRY=0.55; FOXL1=0.53; ZNF354C=0.51; HINFP=0.5;	Irf4, Isgf3g, GCR, PRGR
AGATAT	FOXL1=0.71; GATA3=0.63; NKX3-1=0.61; FOXD1=0.51; ZEB1=0.5;	Gata1, Gata2, Gata3, Gata4, Gata5, Gata6
TTCATA	FOXL1=0.69; NKX3-1=0.62; SPIB=0.61; MZF1_1-4=0.54; CEBPA=0.53; SRY=0.53; CEBPB=0.53; FOXC1=0.52; BATF::JUN=0.52; FOS=0.52; JUND=0.52; ZNF354C=0.51; EHF=0.51; SOX9=0.51; NFIC=0.51;	Pou2f2, Sox12, Sox3, Mtf1, Hbp1, Lef1, PIT1, Sox8, Sox10
TACAAT	FOXC1=0.66; SRY=0.66; SOX9=0.64; FOXL1=0.62; NFIC=0.59; SPIB=0.58; BRCA1=0.58; NKX3-1=0.53; FOXH1=0.53; CDX2=0.52; ZNF354C=0.51;	Sox17, Sox4, Sox5, Sox15, Msx2, Sox13, Sox17, Sox9
CATAAC	BRCA1=0.64; FOXC1=0.61; SRY=0.6; ZNF354C=0.59; FOXL1=0.55; CEBPA=0.54; CEBPB=0.54; BATF::JUN=0.53; MZF1_1-4=0.51; GATA3=0.51; SPIB=0.5; STAT3=0.5;	Sox2, PO5F1, Sox11, CDX1, Foxj1, DBP, HXA10, PO6F1, CDX2
GGCGCC	NFIC=0.62; E2F1=0.54; NRF1=0.52; THAP1=0.51;	Irf4, Isgf3g, GCR, PRGR

Figure 50. Significantly different footprints in primitive with TF predictions. A) Rank analysis to determine the biggest differences in footprint usage between the two cell types, sorted by the most significant footprints in primitive. Red indicates high rank, blue indicates low rank. B) Predictions from JASPAR and FootprintDB databases of potential TFs binding to the identified primitive motifs.

A)



B)

Motif	Jaspar	FootprintDB
ATACCTA	NKX3-1=0.76; FOXL1=0.64; FOXC1=0.61; SPIB=0.57; SOX9=0.55;	ARI5B, NKX31, Zfp187, Pou2f2, Pou2f3, EOMES
TGCGCA	SPIB=0.65; NFIC=0.65; MZF1_1-4=0.62; ZNF354C=0.56; NRF1=0.55; FOXL1=0.54; CEBPB=0.53; CEBPA=0.53; ZEB1=0.52;	E2f family, Max2
ATGCGC	ZNF354C=0.65; SPIB=0.58; FOXL1=0.57; NRF1=0.54; NFIC=0.5;	Nrf1, Gcm1, Gcm2, Zfp57, KLF4, Zfp161, Cebpb, MITF
TACGCG	FOXC1=0.73; MZF1_1-4=0.6; FOXL1=0.58; SPIB=0.56; NFIC=0.54; E2F1=0.52; ZNF354C=0.51;	ARNT, SP100, Rxra, AP2D, Zfp128, KLF7, Zfp128, Max2, Elf5
CGCGTG	BRCA1=0.62; FOXC1=0.62; SREBF1=0.56; NRF1=0.54; SREBF2=0.53; E2F1=0.52; ZNF354C=0.52; MYC::MAX=0.51; MZF1_1-4=0.51;	Zfp161_1, AP2D, Max2, Myf6_2, ATF family
CGTGCG	ZNF354C=0.62; FOXC1=0.61; MAX=0.55; NFIC=0.55; NRF1=0.54; THAP1=0.54; USF1=0.52; SPIB=0.52; ZEB1=0.51; USF2=0.51;	MITF, HIF1A, ARNT, RBPJ, RADR ZIF268, BHLHE40
CATGCG	FOXC1=0.63; BRCA1=0.59; NRF1=0.54; SPIB=0.52; USF2=0.52; USF1=0.51;	Nrf1, Pax2, Gcm1, Pax6, Essra, Pou3f1, Max2, MITF,
ATTATA	FOXL1=0.66; SRY=0.58; NKX3-1=0.56; SOX9=0.55; FOXH1=0.51; CDX2=0.51; GATA3=0.51;	Sox30, Sox14, SRY, Sox21, Tcf7, ZN333
CGCACG	ZNF354C=0.62; FOXC1=0.61; MAX=0.55; NFIC=0.55; NRF1=0.54; THAP1=0.54; USF1=0.52; SPIB=0.52; ZEB1=0.51; USF2=0.51;	MITF, HIF1A, ARNT, RBPJ, RADR ZIF268, BHLHE40
TCGCGG	MZF1_1-4=0.62; NFIC=0.58; E2F1=0.53; E2F4=0.53; FOXC1=0.53; SPIB=0.53; ZNF354C=0.52; E2F6=0.52;	E2f3, HINFP, KAISO, E2f7, Zfp161, E2f2, Max2, E2f5, TYY, MGC1
ATAATA	FOXL1=0.64; GATA3=0.63; NKX3-1=0.61; SOX9=0.57; SRY=0.57; SPIB=0.56; BRCA1=0.56; CEBPA=0.53; CDX2=0.53; FOXH1=0.51;	SRY, Sox30, Pou3f1, Tcf7, Sox14, Glis2, HMGA1, Sox17
CGCGTC	FOXC1=0.56; E2F1=0.52; NRF1=0.5;	Zfp161_1, AP2D, Max2, Myf6_2, ATF family
GCGTGC	BRCA1=0.6; ZNF354C=0.55; THAP1=0.55; NRF1=0.54; ZEB1=0.51;	AHR, ARNT, MITF, RADR ZIF268, DSNR Zfp, HIF1A::ARNT
GTATGT	FOXC1=0.71; FOXL1=0.67; NKX3-1=0.63; ZEB1=0.62; BRCA1=0.6; ZNF354C=0.59; SRY=0.53; GATA3=0.51;	Gm397, Zfp18, Atoh1, DBP, Runx1, Foxd3, CDX1
AATACA	FOXL1=0.69; ZNF354C=0.65; NKX3-1=0.62; SRY=0.62; FOXC1=0.56; BRCA1=0.56; FOXH1=0.55; NFIC=0.54; ZEB1=0.51; SOX9=0.5;	Irx2, Irx3, Irx6, Irx5, ARI5B, FUBP1
CGCCAT	NFIC=0.84; ZNF354C=0.66; MZF1_1-4=0.64; FOXL1=0.63; THAP1=0.63; SREBF1=0.57; SREBF2=0.57; FOXC1=0.55; E2F1=0.53; CEBPB=0.52; CEBPA=0.51; YY1=0.51;	TYY1, Sox14, E2f1, E2f2, Rfx1, E2f4, ZBTB4, TFDP1, TFCREB
TACTAC	FOXC1=0.68; ZNF354C=0.63; NKX3-1=0.62; SPIB=0.58; MZF1_1-4=0.57; FOXL1=0.57;	Hoxb7, Hoxa7, Hoxa9, Mnx1, Hlx9, Zfb187, Spdef
AATAGT	FOXC1=0.58; SOX9=0.58; NKX3-1=0.56; BRCA1=0.54; SRY=0.53; FOXL1=0.51; SPIB=0.5;	Sox1, Sox4, Mef2c, Mef2d, Mef2a, Hoxa3, Zbtb4, Mx2, Nr2e1, Sox17
CCGCGT	FOXC1=0.67; MZF1_1-4=0.62; E2F1=0.53; E2F4=0.53; SPIB=0.53; E2F6=0.53; ZNF354C=0.52; BRCA1=0.51; NRF1=0.51; MYC::MAX=0.51; FOXL1=0.5;	Elf5, Mycn, Zfp161, AP2D, Max2, Mycn, Arnt, Mgc1
ATGGCG	NFIC=0.84; ZNF354C=0.66; MZF1_1-4=0.64; FOXL1=0.63; THAP1=0.63; SREBF1=0.57; SREBF2=0.57; FOXC1=0.55; E2F1=0.53; CEBPB=0.52; CEBPA=0.51; YY1=0.51;	TYY1, Sox14_2, Rfx1, ZBTB4, TFDP1, TFCREB
TCGCGC	MZF1_1-4=0.59; NFIC=0.58; E2F1=0.56; E2F4=0.54; SPIB=0.51; NRF1=0.51;	E2F7, E2F2, E2F5, E2F3, E2F1, E2F4, Zfp161
CGGTGT	MZF1_1-4=0.67; BRCA1=0.67; FOXC1=0.67; ZNF354C=0.6; ELK1=0.54; ZEB1=0.53; SPIB=0.51; FOXL1=0.51;	AR, TBR1, EOMES, Klf1, Tbx5, Pknox2, Meis2, Irf2
AATACT	NKX3-1=0.71; FOXC1=0.59; BRCA1=0.56; SPIB=0.52; FOXL1=0.51; SOX9=0.5;	Ari5b, Irx3, NKX31, Lhx5, Arid5a, Sox30, Ahf, Sox17
GCGTGC	BRCA1=0.57; NRF1=0.54;	NRF1, P53, ZN143, Pou3f1, Osr1, GCM1, P73, Pax2, Ahr
GCGTAC	FOXC1=0.65; ZNF354C=0.62; ZEB1=0.6; SPIB=0.56; FOXD1=0.53; NKX3-1=0.52; FOXL1=0.52; SRY=0.51; HINFP=0.5;	Zfp128, Klf7, EPAS1, Rhox11, HIF1A, AHR, Cebpb
AATACC	NKX3-1=0.57; BRCA1=0.56; REL=0.55; FOXC1=0.53; FOXL1=0.51; NFKB1=0.51; NFIC=0.5; ZEB1=0.5; SOX9=0.5;	Sox3, Ari5b, Sox1, Gabpa, Six2, Gcm1, DDIT3
CGCATG	FOXC1=0.63; BRCA1=0.59; NRF1=0.54; SPIB=0.52; USF2=0.52; USF1=0.51;	Nrf1, Pax2, Gcm1, Pax6, Essra, Pou3f1, Max2, MITF,
CGAGCA	MZF1_1-4=0.7; NFIC=0.61; BRCA1=0.59; ZNF354C=0.57; THAP1=0.54; SPIB=0.53; FOXC1=0.53;	Zfp691, Hmx1, Rfx2, Hmx2, Zfp161, Hoxd13, MITF

Figure 51. Significantly different footprints in definitive cells with TF predictions. A) Rank analysis to determine the biggest differences in footprint usage between the two cell types, sorted by the most significant footprints in definitive. Red indicates high rank, blue indicates low rank. B) Predictions from JASPAR and FootprintDB databases of potential TFs binding to the identified definitive motifs.

We identify 27 and 28 motifs from primitive and definitive cells respectively, that have significantly different average footprints. Top motifs in primitive cells include variations of TATA site (TATATT, TATAAT), and GATA motif. The top motifs in definitive cells include sequences such as ATACTA, and TGCGCA. As a trend, the motifs found to be significantly stronger in primitive cells (red), do not appear to be bound in definitive cells (blue) (Figure 50A). Interestingly, the strongest footprints in definitive cells (i.e. TGCGCA) still have a high score in primitive cells (Figure 51B). Importantly, the difference in score for the same footprint between cell types is higher than would occur by chance. This can only be explained by the fact that a higher scoring motif is one that is bound in more instances in the corresponding cell type.

Some of these motifs are easy to identify, such as the GATA motif, or the TATA site. However, most motifs are either variations of known sites or uncharacterised. In order to comprehensively assess and associate all the differential footprints we found with potential TFs, we curated two open source databases containing TF binding sites: JASPAR¹²⁵ and footprintDB¹²⁶.

(b) From average footprint to protein- TF predictions

The list of top footprints was verified against available TF predicted binding sites from JASPAR and footprintDB (Figure 50B & 51B). JASPAR is the most comprehensive open-access database for TF binding site profiles for multicellular eukaryotes, based on published experiments from various sources and curated by experts¹²⁵. footprintDB is a new TF database that in addition to JASPAR, also

searches across proteomics databases to integrate protein shape and in vitro affinities (UniPROBE) into the annotation¹²⁶.

These databases construct binding site predictions based on position-weight matrices (PWMs). PWMs are the most common algorithms of binding site prediction and they are generated with the assumption that within a set of experimentally identified (ChIP-chip/ChIP-Seq) binding regions, the sequence preferences of a TF are identifiable as compared to background¹²⁷. The output of a PWM is a sequence logo in which the height of the nucleotides is proportional to the likelihood of that nucleotide to be conserved in the motif¹²⁸.

The first observation is the little overlap in the TF prediction results produced by the two databases (Figure 50B & 51B). Also, some of the TF predicted have not been reported in erythroid cells. For example, in primitive erythroid cells, both databases predicted that FOX1L and NKX3.1 bind one of the significant footprints, TATATT. FOX1L is part of the forkhead box proteins, involved in cell growth and proliferation¹²⁹, which was also found to have a pioneer factor function¹³⁰. Considering the other FOX family members in the prediction (FOXD1, FOXC1) as well as the conservation of the binding motifs between the FOX family members in vertebrates^{131,132}, the database results are more likely to indicate that a member of the FOX family is bound at the motif we identified. A possible candidate is FOXO3, considering this TF was previously described to be involved in regulating oxidative stress in erythropoiesis¹³³, and a role in hypoxia signalling was previously described in primitive cells¹³⁴. As for the second common prediction, NKX3.1, even though not commonly described in erythroid studies, it

has been found to be upregulated in primitive erythroid cells¹³⁴, and was described as TAL1 target, in a leukemic cell line¹³⁵.

Even the predictions for the GATA motif bound by the most well-characterised erythroid-specific TF family, the predictions vary tremendously between the two databases and include TFs such as FOX, SOX but also Pou6f1 .

The most plausible explanation for the discrepancies in the predictions lie in the way the databases are constructed: some tissues and cell types are sampled more than others, and certainly cell lines more than primary cells from which experimental genome-wide data is still scarce. In addition, the availability of species-specific PWMs is limited though this is an aspect that the community is working to improve¹²⁵. One other limitation of PWMs, which would affect the predictions, is the way they are calculated. In a PWM, the adjacent nucleotides are considered to be independent, ignoring sequence context. From a cooperative binding perspective, where TF binding is influenced by sequence context and binding partners, this aspect of PWM is a major limitation. Improved algorithms are being developed, such as dinucleotide weight matrices, DWM¹³⁶ or more complex probabilistic models such as the “TF flexible model”¹³⁷, which takes sequence context into consideration. These algorithms have only been tested on a limited number of TFs and they are not yet at a stage where they can be broadly implemented and used, as are PWMs.

As compared to PWMs, which detect occurrence of a sequence, the footprints (digital or average) also quantify binding to that detected sequence. In addition, the footprints do not merge multiple motifs, but rather reveal the preference of binding to any specific instance. This can prove informative as some TFs might

recognise different motifs based on context. And lastly, the footprints are identified and quantified as a unit of 6bp, in which the order of the bases matters. The sequence context is very important for improving the accuracy to detect the bound TFs, and remove ambiguities. Considering 6bp can be short in some instances (some motifs extend beyond 6bp), we are currently developing an option for extending that size to 10-12 nucleotides.

The one aspect that is common between average footprints and PWMs analyses is the fact that neither metric would be able to distinguish between members of a TF family which bind to the same conserved motif (such as GATA1 and GATA2).

(c) Examples of binding locations of differential average footprints

The genomic locations of some of the top average footprints identified in the comparison of primitive and definitive cells were next investigated.

A) Top footprints in Primitive cells

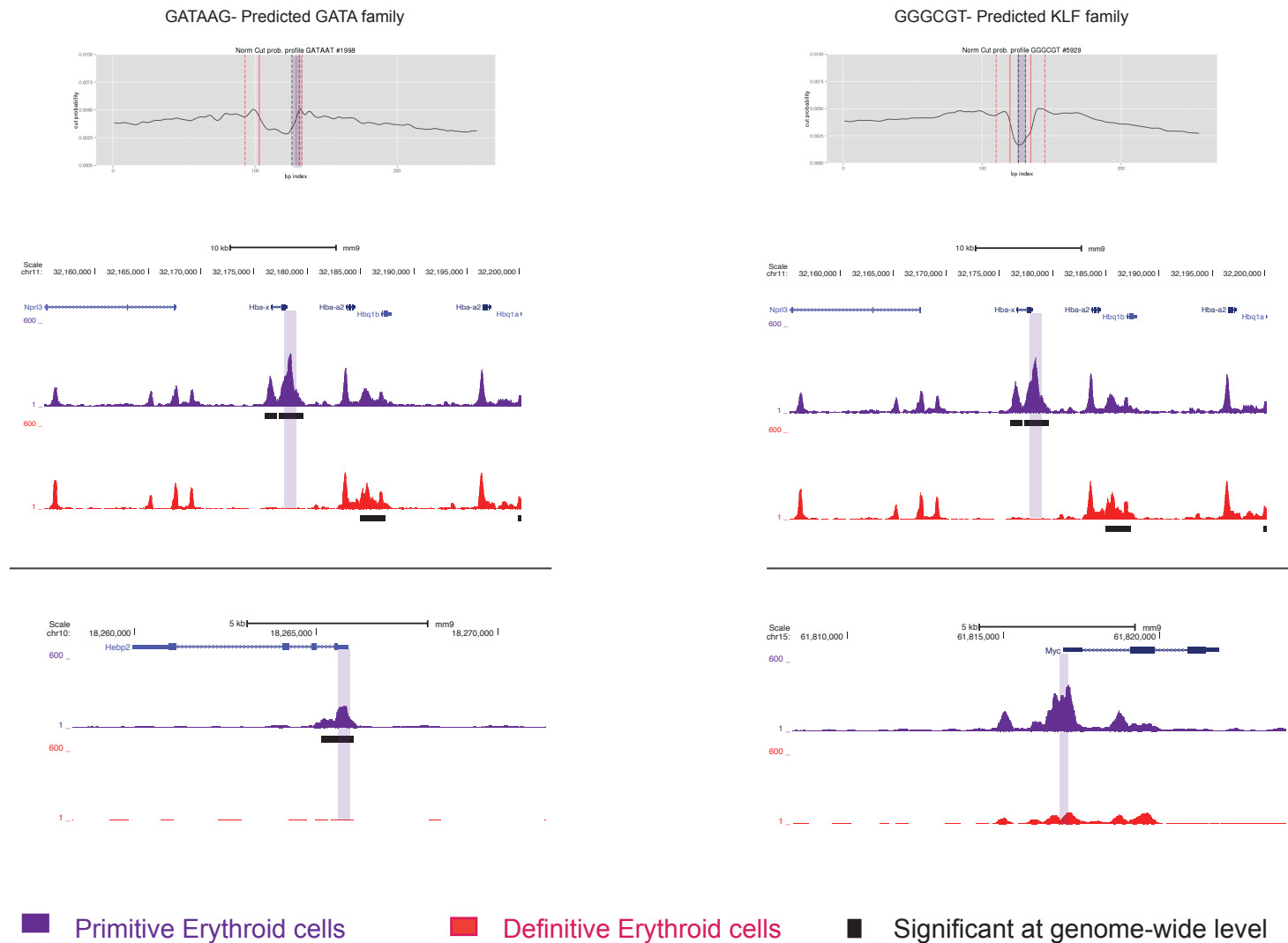
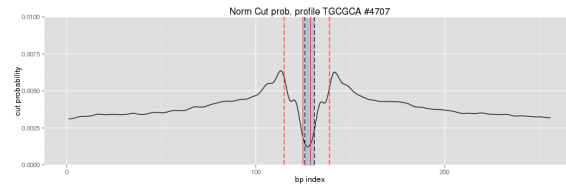


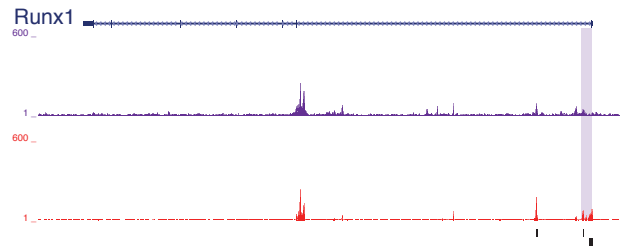
Figure 52. Examples of top footprints in primitive and definitive cells and occurrences in the genome. A) Selected examples of top footprints in primitive cells and genomic regions where these footprints are found. B) Selected examples of top footprints in definitive cells and genomic regions where these footprints are found.

B) Top footprints in Definitive cells

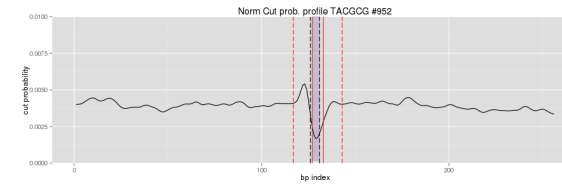
TGCGCA- Predicted SPI-B, MZF1, NRFF1, FOX1L, E2F, MAX2 etc



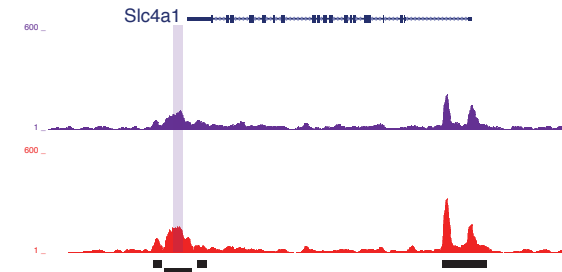
Scale chr16: | 100 kb | 92,650,000 | 92,700,000 | 92,750,000 | mm9 | 92,800,000 |



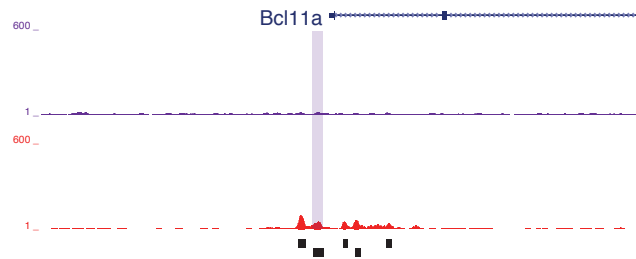
TACGCG- Predicted FOXC1, FOX1L, MZF1, ARNT, KLF7



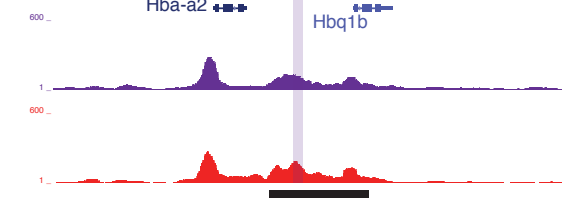
Scale chr11: | 10 kb | 102,210,000 | 102,215,000 | 102,220,000 | mm9 | 102,225,000 | 102,230,000 |



Scale chr11: | 10 kb | 23,965,000 | 23,970,000 | 23,975,000 | mm9 | 23,980,000 | 23,985,000 | 23,990,000 | 23,995,000 |



Scale chr11: | 5 kb | 32,185,000 | mm9 | 32,190,000 |



■ Primitive Erythroid cells

■ Definitive Erythroid cells

■ Significant at genome-wide level

(i) Footprints preferentially bound in primitive cells

In primitive cells, we focused on GATAAG and GGGCGT footprints, ranked at the top of the differential motifs between the two cell types.

GATAAG is predicted to be bound by the GATA family of TF, and is found at 222 primitive enriched regions (out of a total of 1411 significantly enriched primitive regions), majority intergenic and intronic (194), most probably enhancers. We highlighted interesting examples such as the 3' end of the ζ -globin gene, where we were unable to identify footprints previously, or the promoter of *Hebp2* gene (Figure 52A). *Hebp2*, gene involved in heme metabolism, is hypersensitive only in primitive cells. Very little is known about *Hebp2*, except for the fact that it encodes for protein SOUL or p22HBP, and that it binds heme¹³⁸. GATA1 binding has been shown to induce expression of *Hebp1* gene, from the same gene family, in definitive cells⁶², thus it is plausible to assume that GATA1 binding might regulate *Hebp2* as well, in primitive cells. However, we cannot conclude if the binding would result in induction or repression of the gene.

GGGCGT is predicted to be bound by members of the KLF family by both databases. We find this footprint in 189 significantly enriched primitive peaks (76 intergenic and intronic and 73 at promoters). We find this footprint at the 3' end of the ζ -globin gene, making KLF1/2 plausible candidates as they have been described to be involved in regulation of the embryonic β -globin genes¹³⁹. Another highlighted region is the Myc promoter, shown by the same study¹³⁹ to be bound and regulated by KLF1/2 (Figure 52A). These binding events in primitive cells would need to be confirmed with further studies such as ChIP-Seq.

In addition, we also found this footprint in definitive cells, downstream of *Bcl11a* promoter which is in agreement with studies showing KLF1 regulates γ -globin silencing by up regulation of Bcl11a¹⁴⁰.

The fact that the footprint does not coincide with the consensus CCAC-motif, but with the reverse complement that is found at validated regions, supports the need of integration of 'motif grammar' such as orientation and spacing, in our studies of regulatory sequences, as it may play an important role in understanding TF cooperativity and interactions.

(ii) Footprints preferentially bound in definitive cells

For TGCGCA, one of the top footprints in definitive cells, there appears to be no consensus between the database predictions in terms of the TFs which bind this motif. Candidates include SPI-B, MZF1, NRF1, C/EBP α , C/EBP β , E2F family, Max family. This footprint occurs at 135 significantly enriched definitive regions (out of the total of 2676 significantly enriched definitive regions). Solely from the footprints, it is difficult to identify which of these factors, or how many bind the footprint. Both E2F-2¹⁴¹ and MAX have been reported to be involved in terminal erythroid differentiation¹⁴², and thus are potential candidates. Even though we were limited in terms of identifying the TFs, we explored the locations of this footprint (Figure 52B). Binding sites include the proximal regions of the promoters of Bcl11a and Runx1.

TACGCG, another significant footprint in definitive does not reach a consensus in terms of TF predictions. Examples of regions where we find this footprint are the

promoter of Slc4a1, a band 3 anion transport protein as well as downstream of the α -globin promoter, in the hypersensitive region of unknown function.

4.03 Computing “average” footprints from ATAC-Seq data

Having developed an unbiased way to interrogate DNase-Seq data in the form of average footprints, we next wanted to test if we can apply this approach to ATAC-Seq data. The benefits of being able to identify all the bound sequences in ATAC-Seq data are obvious: we would be able to detect binding at high resolution from input material in a faster manner. ATAC-Seq average footprints would allow us to analyse the differential binding at the two different stages of erythroid differentiation.

As discussed in Chapter 1, even though we can reliably and reproducibly enrich for open chromatin regions using ATAC-Seq, we could not detect digital footprints when analysing a single locus. The average footprints approach offers the advantage of analysing genome-wide occurrences of motifs, thus highlighting the existence of a pattern which might not be visible at a single region in the genome.

In order to test and compare if ATAC-Seq footprints are similar to DNase-Seq ones, we applied the average footprints method to DNase-Seq and ATAC-Seq from the same cell type, definitive erythroid cells, and at the same genomic regions. If the methods detected the same type of binding, suggested by the similarity in identifying open chromatin regions, we would expect to see similar motifs having strong footprints. In order to assess any potential sequence bias, and to improve normalisation, we have also generated control tracks for both assays by performing DNaseI and ATAC on genomic DNA, rather than chromatin.

(a) Average footprints generated from ATAC-Seq show a preference for A/T sequences

All 4096 potential motifs were tested and their corresponding footprints were generated and sorted and ranked. The most significant differences in motif usage were calculated using the same rank product analysis as described in the section above. We identify 98 motifs to have significantly different footprints between the two methods, in the same cell type.

When sorting the motifs by the strongest footprints in DNase-Seq, we noted only a small degree of overlap: the strongest footprints (i.e. CACGTG, TGACGT etc) are not identified as strongest footprints in ATAC-Seq, having significantly lower score (colour yellow, orange) (Figure 53). However, the more striking discrepancy was visible when assessing the strongest footprints in ATAC-Seq: not only do the top ATAC-Seq motifs have extremely low scores in DNase-Seq (colour blue) suggesting they do not form footprints, but they are also enriched for the A and T bases.

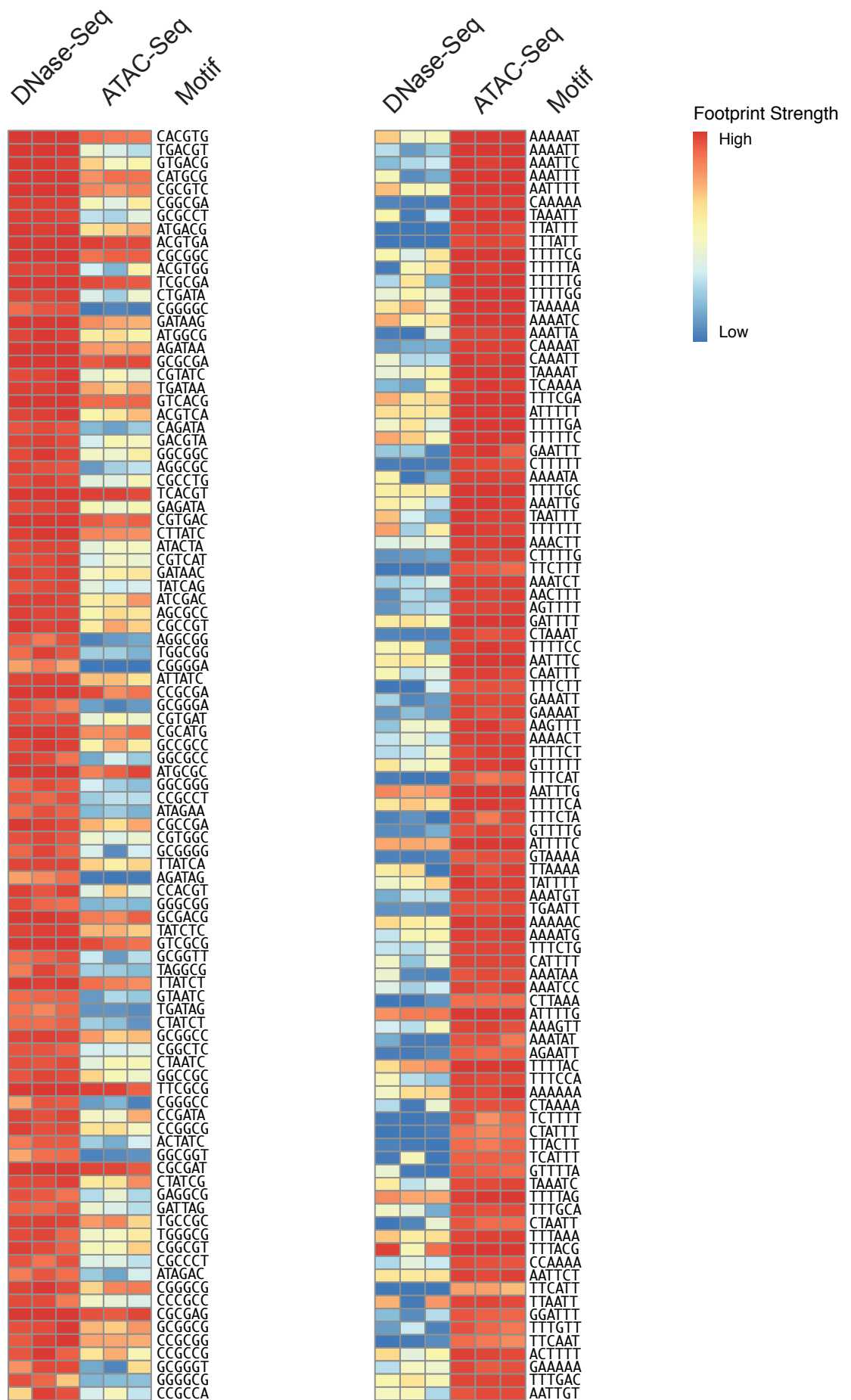


Figure 53. Comparison of the strongest footprints in DNase-Seq and ATAC-Seq in definitive erythroid cells. Rank analysis to determine the biggest differences in footprints identified with the two different methods. Red indicates high footprint score, blue indicates low footprint score. Footprints were generated in the same cell type, and within the same genomic regions. Shown are the motifs corresponding to the strongest footprints sorted by the DNase-Seq (left) and ATAC-Seq (right).

We hypothesised that a potential sequence bias would be visible in the control (background) dataset for ATAC-Seq. We generated the average footprints for all the motifs in the control datasets to compare (Figure 54).

The motifs with the strongest footprint in ATAC-Seq, 'AAAATT' and 'AAATTT' show no footprint in the DNase-Seq. The plots from background data are expected to be noisy, with no clear pattern, as can be seen in the DNase-Seq background plots. In contrast, the control plots from ATAC-Seq show an unexpected similarity to the test dataset. All the top sequences we tested showed a mirroring pattern in the control data set suggesting the Tn5 transposase probably has a sequence preference for A/T rich regions.

Next, two top motifs from DNase-Seq, the E-box 'CACGTG' and GATA site "GATAAG" were tested in ATAC-Seq data. Recognisable footprint profiles are generated from the ATAC-Seq, with a few differences: first, the ATAC-Seq footprints do not have the same depth (judged by the score and pattern) as the DNase-Seq footprints; second, the palindromic E-box motif forms a stronger footprint, than the non-palindromic GATA; third, the footprints are wider in ATAC-Seq as judged by the detection of the footprint shoulder. The control tracks show no reminiscent pattern of these footprints.

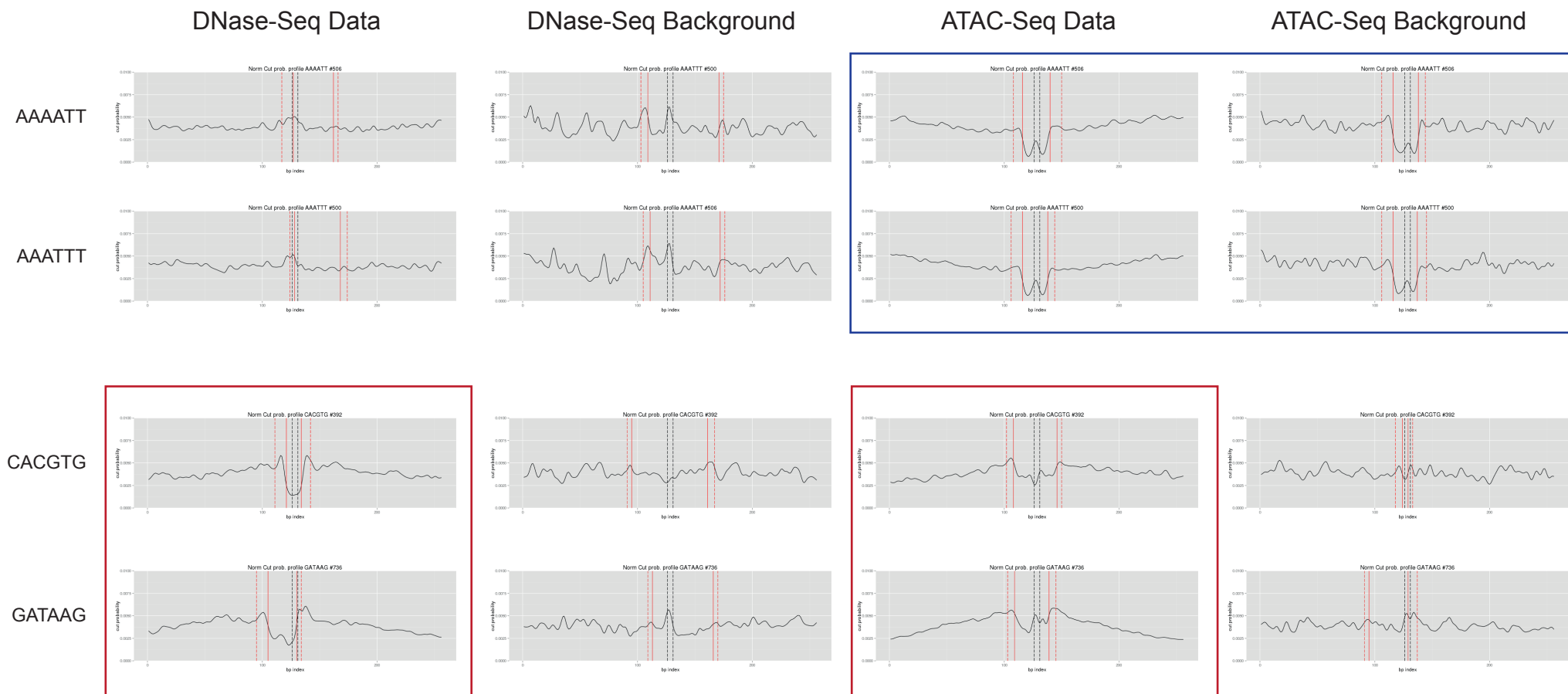


Figure 54. Profiles of the motifs with the strongest footprints as identified in the same cell type by DNase-Seq and ATAC-Seq. The motifs with strongest footprints in DNase-Seq and ATAC-Seq and profiles for the same motifs in the corresponding background data sets. Background data was generated by performing the DNaseI or ATAC assays on genomic DNA rather than chromatin. The processing of the background data sets is identical to the test data sets.

Analysis of genome-wide average footprints from ATAC-Seq as compared to DNase-Seq shows that the two methods enrich for open chromatin fragments through mechanisms that affects TF binding to different extents. Transposase-mediated cleavage occurs with bias which is detectable when comparing genome-wide average footprints generated with the two assays, as different motifs appear to be bound to significantly different degrees. The potential biological causes of the bias and variability, followed by technical causes will be addressed in the Discussion-Open-chromatin assays.

The aim of the DNase and ATAC average footprints comparison was to validate the use of ATAC data with this novel approach, so that we can apply it to identify the sequence usage in the differentiating populations for which only ATAC-Seq data is available (Results Chapter3). Considering the bias identified in ATAC-Seq, we conclude this aspect needs further investigation in term of transposase mechanism and appropriate modelling for computational analysis. Thus, we could not pursue the high-resolution analysis of sequence usage in early progenitors and differentiating cells.

4.04 Conclusion

In this chapter we presented a novel, unbiased approach of “average footprints” that can be applied to identify patterns of sequence usage in different cell types. Binding of different regulatory sequences at HS regions is reflected in the strength of the DNase average footprints and can be used to rank more frequently used motifs within regulatory sequences. We identify “primitive – specific” and “definitive-specific” average footprints and the corresponding motifs. Based on the information available in databases such as JASPAR or footprintDB, potential TFs binding to those sequences can be predicted. We believe this workflow should serve as an indication of the families of TFs involved. In order to identify the particular TFs that drive cell specific differences the average footprints can be used in combination with other methods (ChIP-Seq for specific protein binding, RNA-Seq for expression).

ATAC-Seq data was tested with the same approach and the “average footprints” revealed sequence bias for A/T regions. In addition, the average footprints do not solve the resolution issue of ATAC-Seq we observed in Chapter 1: average footprints are detected but they are wider and less strong than DNase footprints. We conclude the transposase mechanism needs further investigation for understanding how to interpret high-resolution data.

The average footprint approach can also be applied to investigate changes of usage of specific regions of interest. We adapt this novel approach to allows us to investigate the role of sequence variation in gene regulation.

Chapter V. Analysis of variants in-regulatory sequences in human erythroid cells

5.01 Introduction

The findings that variants associated with common disease are enriched in the noncoding genome ⁴⁶, and that they affect gene expression in a tissue- and allele- specific manner^{143,144} were indicative of the fact that GWAS SNPs concentrate in *cis*-regulatory elements ¹¹.

Even though the mechanism by which these variants exert their effect is yet to be determined, their location in regulatory regions suggests they potentially impair TF binding, and affect enhancer function.

In order to analyse the function or dysfunction of an enhancer, it must be linked to its target gene so the transcriptional output can be measured. This is far from trivial as enhancers tend to be located distal from their target genes, at varying distances, and can regulate multiple genes. It is no surprise this is one of the main challenges in gene regulation: linking enhancers to their target genes, genome wide. This is where model loci come into play, as their extensive characterisation most often covers functional annotation of their corresponding regulatory elements.

Evidently, the extensively characterised human α -globin locus provides an excellent model for investigating the role of sequence variation in disease. Not only the enhancers have been functionally linked to their target α -globin gene,

there are also naturally occurring variations in these enhancers, of interest the SNPs within the major regulatory element R2/HS-40.

The human α -globin cluster is located at the telomeric end of chromosome 16 and is regulated by four highly conserved erythroid enhancers (R1 to R4), R1 to R3 located in the upstream *Nprl3* gene. Although all 4 elements fulfil the characteristics of enhancer sequences (conserved, marked by histone modifications and bound by TFs), functionally, R2/HS-40 has been shown to play a critical role in regulating the cluster (Introduction- Variation in regulatory sequences) and will be the focus of our investigation.

R2/HS-40 is the most highly conserved of the four enhancers in the 22 species studied⁵¹. The multispecies alignment reveals a core of binding motifs (~90bp) containing a GATA binding site (CAGATAAC), two AP1/NF-E2 binding sites (CATGACTCAG.....TGCTGAGTCAT), and a second variant GATA site (GCTGATTA). The order, the sequences and the spacing between these motifs is conserved⁵¹. The HS region also contains other, less conserved motifs (GATA, CACC). These consensus sites have been shown to bind TFs *in vitro*, and, by footprinting and *in vivo*, by ChIP experiments^{41,42,145,146}.

In this chapter we describe an adaptation of DNase “average footprints” (Results-Chapter 4) for predicting the effect of variants on protein binding. Using this adaptation, we investigate variants in the α -globin major regulatory element and their potential effect on the TF binding sites. We first perform DNase-Seq in primary human erythroid cells expanded and differentiated in culture (Materials and Methods) then generate the DNase-Seq digital footprints (see Results-

Chapter 1) specifically at the human α -globin locus. Sequences identified by these digital footprints are then used to derive “average footprints” for human erythroid cells (Results-Chapter 4). We then calculate “damage scores”, (a prediction of the effect of a base change on the formation of an “average footprint”) for the variants in R2/HS-40. Finally, we present the case of a patient with unexplained anaemia, who harbours one of these variants in R2 and test if this SNP could affect the enhancer function and explain his condition.

5.02 HS regions at the human α -globin cluster

DNase-Seq was performed on cultured human erythroid cells. As expected, hypersensitivity highlights the human α -globin promoters and the four regulatory elements (Figure 55). The HS regions are marked by chromatin marks (H3K4me1, and H3K4me3), PolIII and TFs such as SCL (DeGobbi 2007, Lower 2013, and unpublished data). The high level of hypersensitivity of R2/HS-40, relative to the other elements in the cluster is clearly indicative of its role in the transcription of the α -globin gene.

We next generated digital footprints from DNase-Seq (Introduction/Chapter1) in order to visualise and assess the sequences directly bound at the HS regions in the human α -globin cluster.

(a) Digital footprinting in human primary erythroid cells

Our data shows for the first time DNase-Seq digital footprinting in primary human erythroid cells (Figure 56).

We assessed the presence of previously described highly conserved regions in the digital footprints of the elements in the cluster. We show the profiles of R1, R2 and the α -globin promoter. In R1, we observe that the two conserved GATA1 sites previously described⁵¹ do not form a clear footprint; however, a species-specific site is visible and corresponds to a variation of the Maf-element known to dimerise with the erythroid specific TF, NF-E2. In R2/HS-40, the previously described AP1/NF-E2 sites are clearly picked as seen on the track, as well as one of the conserved GATA1 motifs. As for the promoter region, we can observe a footprint at the conserved GATA1 site, as well as the conserved CCAT motif.

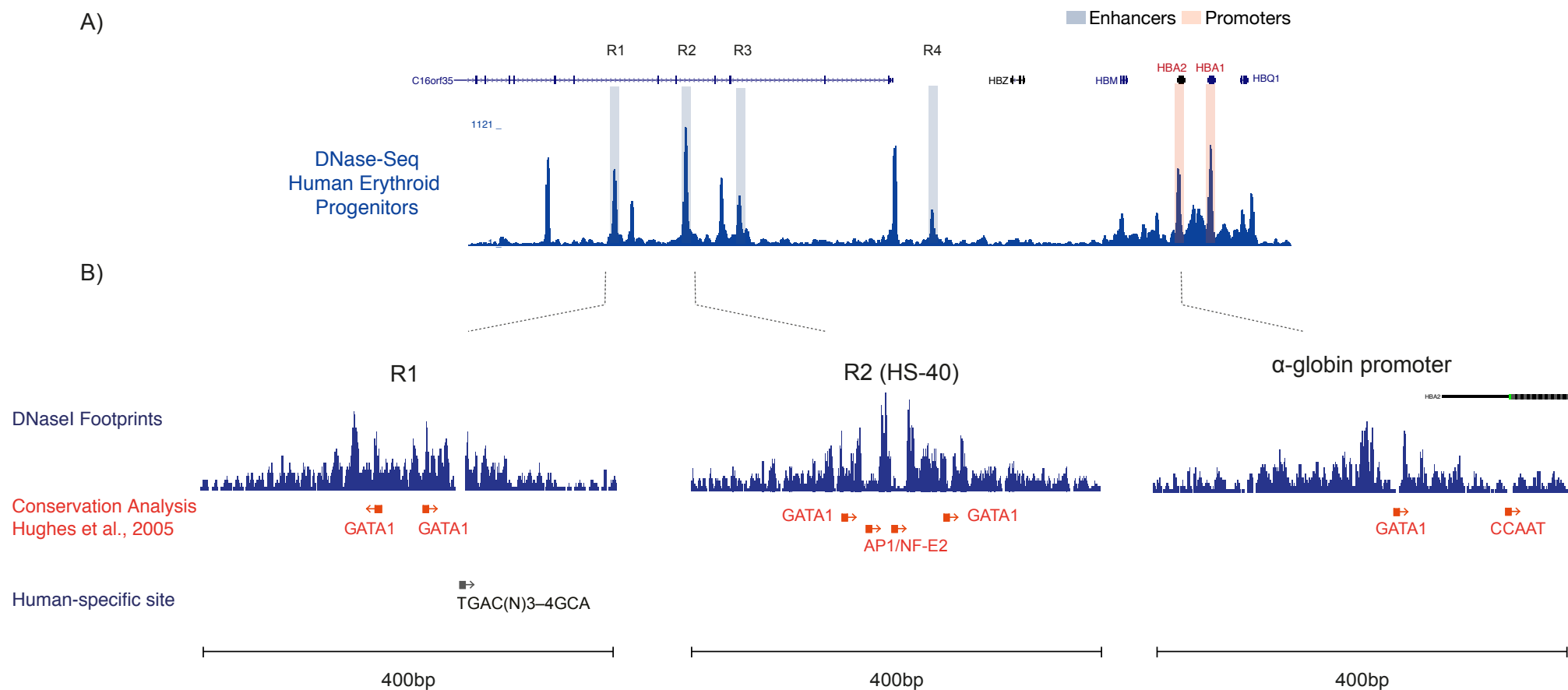


Figure 56. DNase-Seq and DNase Footprints at the human α -globin cluster . UCSC tracks showing HSs at the α -globin cluster. DNase-Seq from human progenitors and DNase footprints at R1, R2 (HS-40) and α -globin promoter.

(b) Digital footprinting of the major regulatory element R2 (HS-40)

We compared our predictions of protein-binding based on footprinting data with previous studies on protein occupancy at R2/HS-40- either in vitro ⁴¹, in vivo by DMS footprinting⁴² and site-directed mutagenesis¹⁴⁵, or ChIP-qPCR /ChIP-seq^{146,147}(Figure 57). Interestingly, the four highly conserved motifs are identified as bound in both our analysis and the previous reports. The motifs show protection from DNaseI cleavage (highlighted), though the strength of the digital footprints is not the same for all four motifs. The two AP1/NF-E2 sites have clear, well-defined footprints, followed by the GATA1 site downstream. The GATA1 site upstream of the AP1 motifs shows a weak footprint.

In addition to the conserved sites, the previous in vitro and in vivo DMS footprinting studies identified additional bound GATA1 and CACC sites, which are located outside the syntenic region. These two additional GATA1 sites are not clearly identifiable in our digital footprinting data whereas one of the CACC-box sites (identified in vitro ⁴¹) is detected by the footprint caller in our data.

The contribution of these sites to the transcription of both the α -globin and ζ -globin genes was investigated by site-directed mutagenesis in human erythroid K562 cell line and mouse erythroleukemia (MEL) cells. The two AP1/NF-E2 sites as well as the CACC-box and GATA1 sites downstream of the AP1 motifs were found to contribute either positively or negatively to α -globin expression, in one or both of the cell lines ¹⁴⁵.

Even though there is consensus between previous in vitro studies and digital footprinting on which are bound by TFs at R2/HS-40, there is only partial agreement on the functional significance of all these bound sites as shown by site-directed mutagenesis in cell lines with transiently expressed constructs. It is thus not clear, which of the binding sites are necessary for the activity of α -globin activity in vivo, and which, if mutated could have a detrimental effect on gene expression.

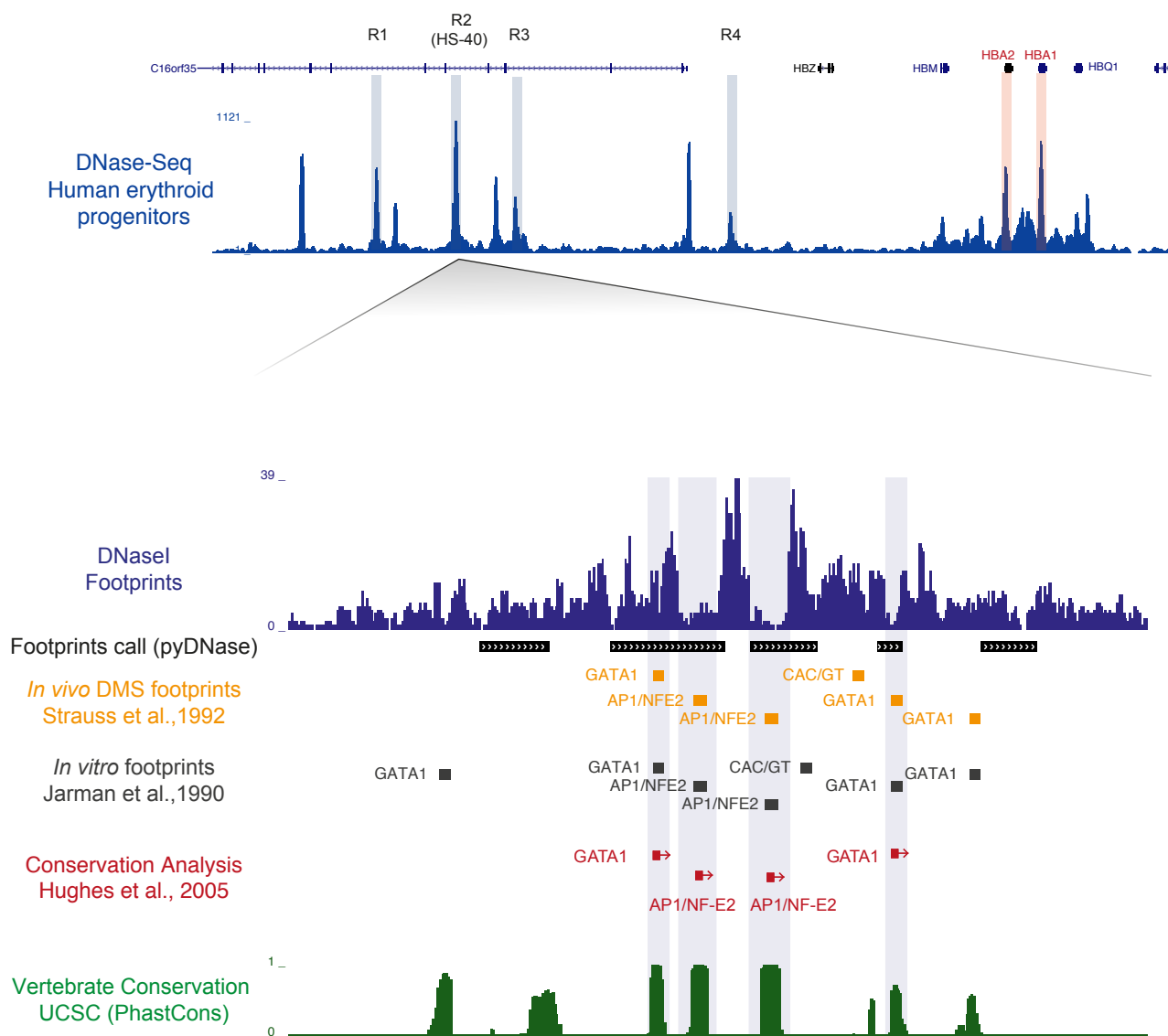


Figure 57. Footprints at the major regulatory element of human α -globin. UCSC tracks showing HSs at the α -globin cluster. DNaseI footprints compared to previously identified in vitro and in vivo DMS footprints at the major regulatory element of the human α -globin cluster.

(c) Sequence polymorphisms in R2/HS-40

The study of the α -thalassemic phenotype in patients has revealed that, in addition to deletions of the upstream elements, some patients present point mutations in R2/HS-40¹⁴⁸ (unpublished data from the lab).

We find several variants located within the R2/HS-40 site. We have evidence that a variant can impact on *cis*-regulation as was previously reported by us^{122,147,149} and others¹⁵⁰⁻¹⁵². In order to prioritise which of the 12 variants located in R2 is to be functionally studied, we adapt the “average footprints” analysis to predict potential effect of the SNP on the formation of footprint (in other words protein binding) through calculation of variants “damage scores”.

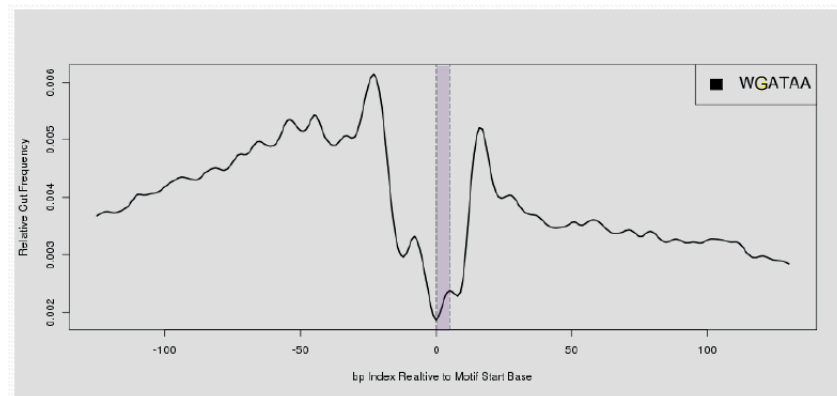
5.03 Using “average” footprints to assess impact of variants in regulatory sequences

(a) Assessing effect of one variant on a sequence.

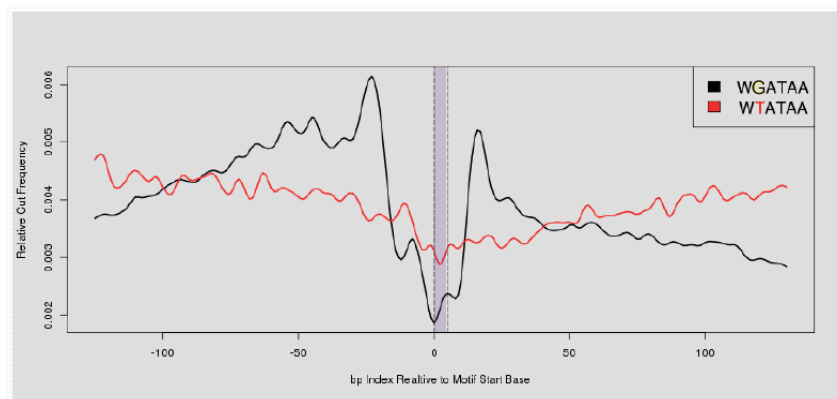
The new approach to generate “average footprints” described in Results-Chapter 4, was adapted to investigate the effect of sequence variation on the formation of footprints. Thus, by changing the sequence of a known binding motif (referred to as the reference), we can test if the new, altered sequence (the variant) forms a footprint in the cell type investigated.

The difference between the reference footprint and the variant footprint can be calculated in the form of a “damage score”. This allows for the use of the “average footprints” as a method to predict alteration of a binding site (Figure 58).

Reference



Variant



Variant

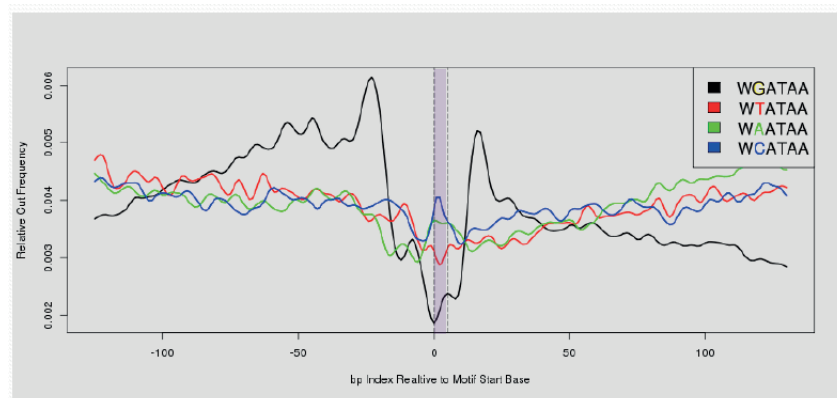


Figure 58. Effect of sequence alteration on average footprints. Examples from mouse DNaseI data where the WGATAA motif was altered by replacing the G with T (red), A (blue), C (green). The location of the WGATAA motif is highlighted. Plot generated in collaboration with Ron Schwessinger.

(b) Assessing damaging effect of changes across an entire region

This method can be applied across an entire genomic region, to identify sequences sensitive to variation.

One nucleotide can be part of multiple 6bp motifs, and occupy different positions (1-6) if one imagines a sliding window of one base pair over the sequence of interest. For each of these positions in a motif, we calculate the “reference footprint”. We then change the nucleotide analysed with one of the other 3 possible bases, and generate the “variant footprint” at each position in the motif, over an entire sequence of choice. Finally, we calculate the “damage score” (difference between the “reference footprint” and “variant footprint”) for each position the nucleotide occupies in the motif, and plot the highest “damage score”.

Thus, by changing each nucleotide in the sequence we are interested in, we can assess if the “variant footprint” still forms an average footprint and calculate the difference with respect to the “reference footprint”. We calculate the “damage score” in this manner for each bp change in the sequence. The higher the “damage score” associated with a bp change, the higher the likelihood that this variation would alter the binding site. A “damage score” lower than 0 would indicate formation of a new binding site.

We assessed in this manner each nucleotide, as part of a 6bp motif, in R2/HS-40. The resulting graph displays the “damage scores” of the “average footprint” generated as described above (Figure 59).

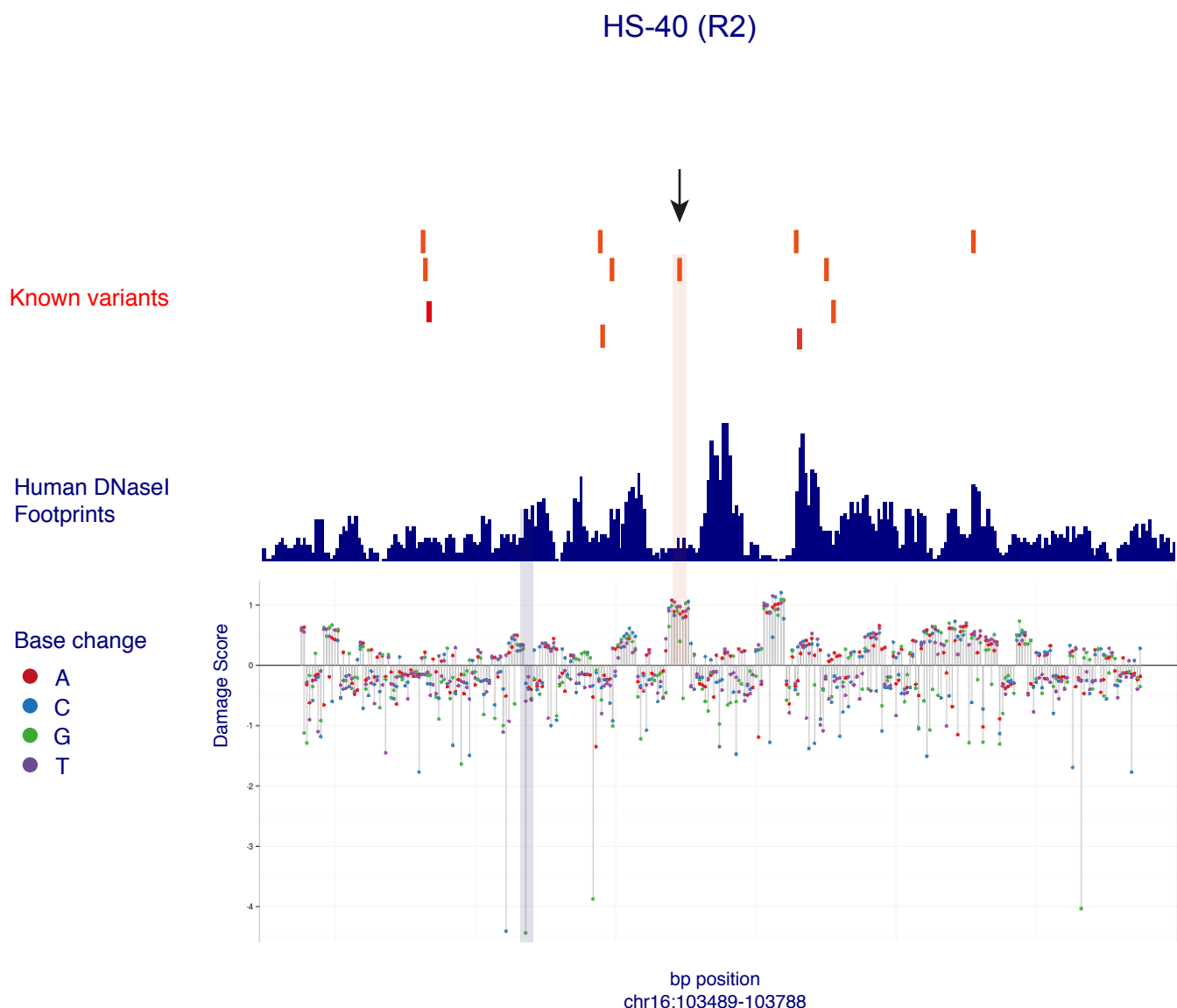


Figure 59. Calculated damage scores of sequence variations at the HS-40. DNaseI footprints shown in blue. The damage score plot shows the predicted damage of footprints caused by changing each base pair. Known rare and common variants shown on top. Highlighted in red is a variant with one of the highest damage scores; in green the common variants identified by GWAS (Harst, 2012); in blue, the region with the lowest negative damage score, predicted to form a new footprint. Plot generated in collaboration with Ron Schwessinger.

The highest damage to a footprint is predicted to occur if variation is introduced in the highly conserved AP1/NF-E2 sites. The conserved GATA1 sites, as well as the GATA1 site upstream of AP1/NF-E2 sites (identified by in vitro studies) also had increased damage scores.

In order to see if any reported variants overlap with these damage-prone regions, we mapped the known sequence R2/HS-40 variants along with the “damage scores”. One of the reported rare variants from a patient with unexplained anaemia (rs177510), overlaps with one of the conserved AP1/NF-E2 site, and has high “damage score”.

In addition, there are four base-pair changes with extremely high “damage scores” with a negative value, in regions that do not form footprints in the reference suggesting such changes could form new binding sites. No variants were reported in these regions.

We decided to investigate the most interesting observations further: the damaging variant within the AP1/NFE2 site, and the variant with the highest prediction to form a new binding site.

(c) Variant predicted to damage a binding site

(i) Patient with unexplained anaemia has variant in AP1/NF-E2 site

We selected patients from a library of referrals to the Higgs' laboratory for investigation because of unexplained anaemia. No mutations affecting the promoters are found but variants in HS-40 were detected in the patients' α -globin locus. Two cases have the same variant, rs177510, in the highly conserved AP1/NF-E2 site.

Patient RE has agreed to be part of our study, thus, will be the focus of our further investigation. The patient's haemoglobin and red cell parameters were measured to confirm the initial anaemic phenotype. Haematological indices are lower than normal (reduced hemoglobin-Hb, mean cell volume- MCV and mean cell hemoglobin-MCH), and suggest a mild anaemia (Figure 60). RE has never required transfusions and is otherwise well.

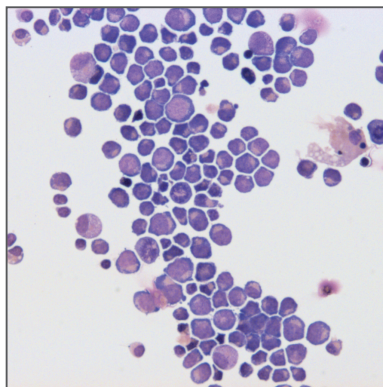
A)

	Patient RE	Normal Range
Hb [g/dl]	10.8	12-15
MCV [fl]	80.1	83-105
MCH [pg]	23.7	27-32
RBC [count/l]	4.56×10^{12}	$3.8-4.8 \times 10^{12}$

B)

RE sample

Day 8



Day 10

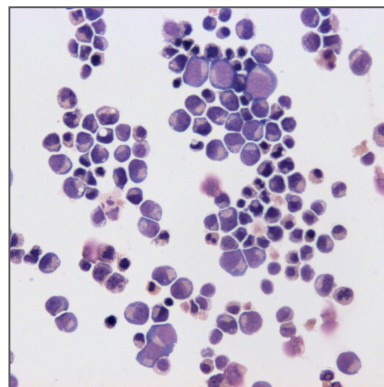


Figure 60. Characterisation of patient's RE blood sample. A) Haematological indices for patient RE, as compared to the normal range. B) Giemsa-stained cytopspin preparations from patient sample Fibach cultures.

(ii) Predicted effect of variant rs177510

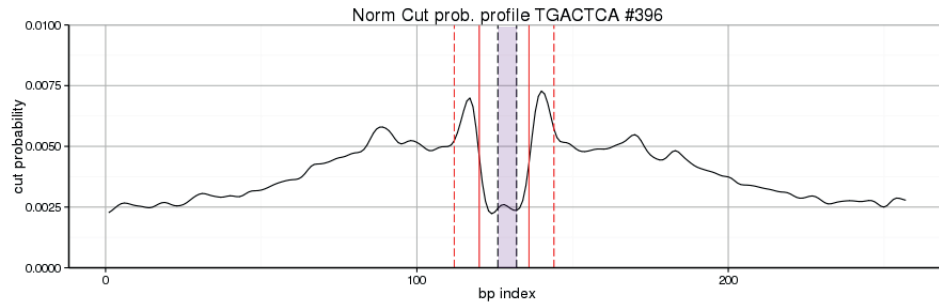
As reported above, the rs177510 variant has a high footprint “damage score” (Figure 59) and is predicted to affect the binding site in which it is located. In order to visualise this particular binding site and the predicted effect of the variant, we generated the “average footprints” for the reference version of the AP1/NF-E2 site found in normal individuals, as well as the “average footprint” of the sequence variant found in the patient (Figure 61). There is a marked difference in the two average footprints: the reference shows the clear contrast between the bound site and the surrounding highly cut regions; whereas the variant shows no clear footprint and high background, suggesting that the new sequence is not bound (as detected by DNaseI footprints) in human erythroid cells.

Based on the anaemic phenotype and the prediction of the damaging effect in the binding site, we decided to pursue the investigation and determine if the change in R2 regulatory sequence can explain the reduction in the α -globin gene expression.

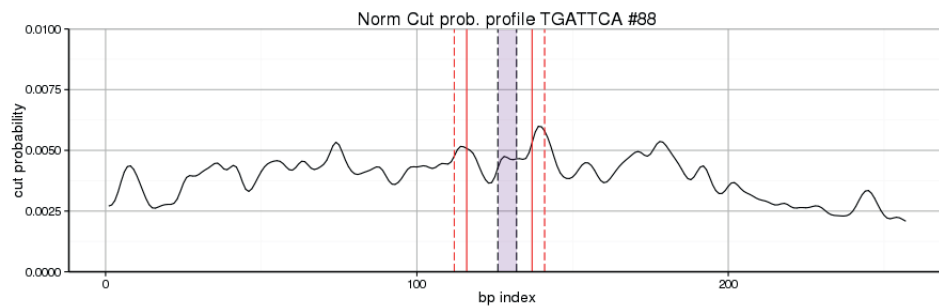
However, due to time constraints, only limited amount of functional validation could be carried out for the thesis. In addition to preliminary work presented below, future plans on how this hypothesis will be investigated as part of my postdoctoral project are described in the last part of the Discussion.

A)

Reference



Variant



B)

NF-E2 binding motif in HS-40:



Variant: C> T

Figure 61. Effect of sequence variant on the NF-E2 footprint. A) The average profile of the AP1/NF-E2 motif found in HS-40 (Reference) and the average profile of the motif containing the variant rs177510 (Variant). B) NF-E2 binds an AP1 motif in HS-40; the variant introduces a change from a C to a T.

(iii) Preliminary work for functional validation

Primary erythroid progenitors from patient RE were expanded and differentiated in order to analyse the open-chromatin landscape at the α -globin cluster.

The differentiation of the progenitors in Fibach culture system (Materials and Methods) progressed normally as assessed by morphology (Figure 60A) and immunophenotyping by FACS (data not shown). Thus, the effect the variant might have on R2 enhancer function does not seem to affect erythroid differentiation, crucial information that allowed the differentiated population from patient sample to be compared to that of normal controls (no population bias).

Due to the faster processing time, we prepared ATAC-Seq libraries for an initial assessment of the HS regions in patient erythroid cells. In the near future, the DNase-Seq footprints will be generated for a high-resolution analysis of the AP1/NF-E2 binding site (Chapter1- Digital Footprinting).

The analysis of open chromatin using ATAC-Seq reveals that all the HS regions observed in the control sample, are present in the patient RE, including R2/HS-40 (Figure 62). The level of enrichment at R2/HS-40 is the highest relative to the other regulatory elements in the cluster, in both the control and the patient sample.

The HS regions neighbouring the four enhancers corresponding to CTCF sites and the *NPRL3* promoter have reduced hypersensitivity when compared to the same regions in the control sample. Considering the control and the patient samples, have been processed with the two different open-chromatin assays

(DNase-Seq and ATAC-Seq respectively), these variations in enrichment can only be validated once we obtain the DNase-Seq profile from the patient data.

The potential effect of the variant is not detected at low resolution, when ATAC-Seq hypersensitivity track is examined at the R2 enhancer. This result does not exclude an effect of the SNP on the NF-E2 binding site. However, this could only be evaluated upon performing DNase-Seq followed by digital footprints analysis at R2/HS-40. Until this is completed, we cannot draw any further conclusions.

(d) Variants predicted to form a new binding site

Amongst the base pair changes with potentially interesting effect on R2/HS-40, we noted variants with damage scores indicative of formation rather than disruption of binding sites. This aspect of the prediction approach is novel and will have a valuable contribution to the field of GWAS bioinformatics analysis, as predicting formation of new binding sites within regulatory sequences is extremely challenging, if at all possible, with the published and currently used methods.

(i) Formation of new footprints, at sites predicted in silico and validated sites

With the aim of investigating the formation of a new binding site, we analysed the motif containing a base change with high “damage score” with a negative value in R2/HS-40 (Figure 63).

The sequence variation is a change from a C to a G: CAG**C**TG to CAG**G**TG. Even though both motifs are variations of the E-box, [CANNTG], only the latter forms a footprint in the human erythroid progenitor data.

CAGCTG

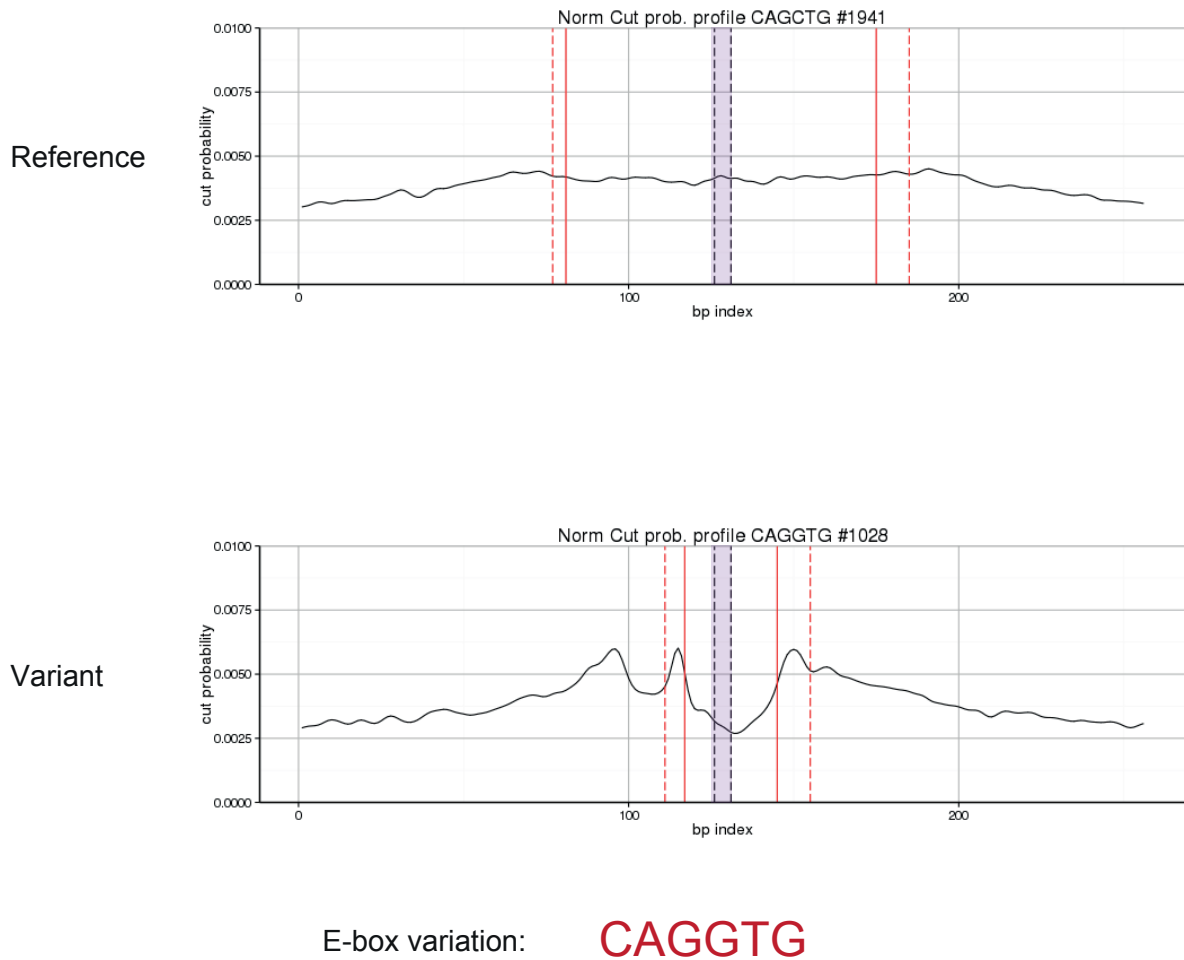


Figure 63. Predicted new footprint in HS-40. The sequence variation predicted to have the lowest negative damage score could form a new footprint. Average footprint of the reference sequence extracted from HS-40 and the footprint of the sequence with a basepair change.

It is not surprising that the variant sequence has such a strong footprint as the CAGGTG E-box is common motif in erythroid regulatory regions bound by erythroid TFs such as SCL/TAL1, or E2A^{61,72,153}. Interestingly, when analysing the “average footprint” of this variant motif, we observe a second protected site upstream of the E-box (Figure 63). This is very likely to be the GATA1 site, that occurs as part of the extended motif of [GATA(9-10)CANNTG] reported to be bound by the pentameric complex⁷² within erythroid regulatory regions^{71,154,155}. This finding highlights the power of this tool in detecting the “average footprint” not only at the motif or variant under investigation but also at neighbouring sequences if they occur simultaneously at detectable frequency. The result of this analysis also emphasises the need for developing the option of extending the motif search beyond the present option of 6bp, aspect we are currently working on.

The biological significance of a predicted detrimental variant could be scrutinised further if the SNP is introduced in the “normal” sequence in primary erythroid cells (by genome editing) followed by analysis of protein binding and α -globin expression in culture.

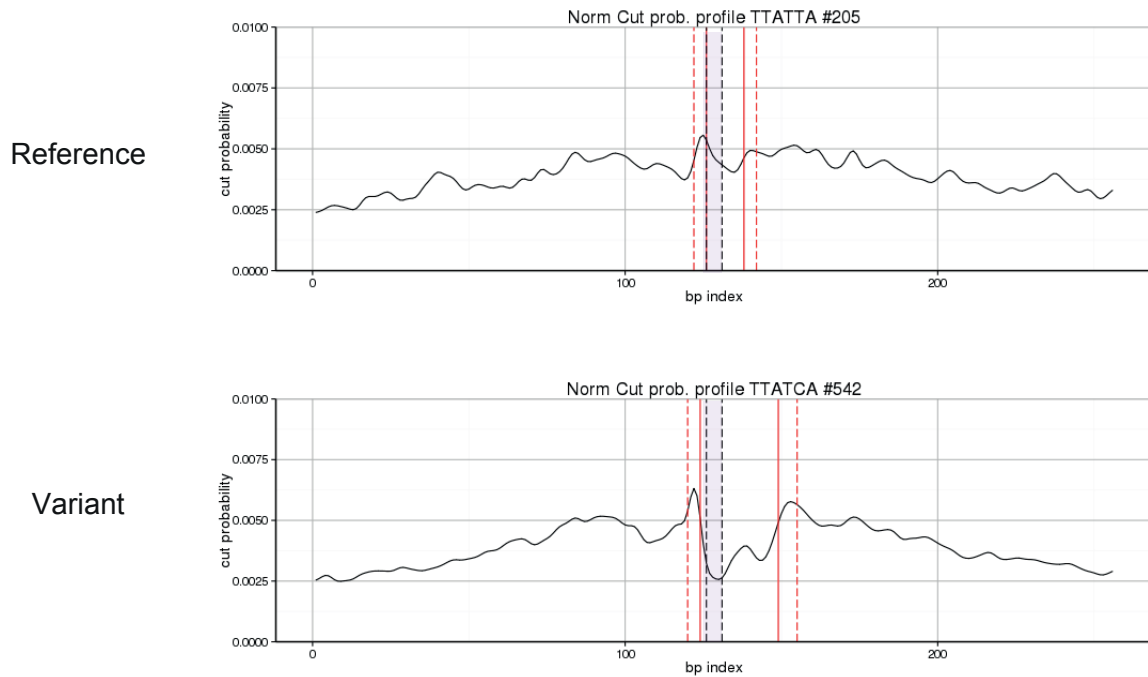
In order to validate the prediction approach for formation of binding sites, we tested a published SNP referred to as the Vanuatu mutation that has been associated with a functional effect on α -globin expression. The mutation introduces a change from a T to C at position 149709 (build hg18) upstream of the α -globin promoter, creating a GATA1 site on the reverse strand that leads to the formation of an alternative promoter that results in significant reduction of α -

globin expression, presumably by diverting the enhancer function away from their target α -globin promoter¹⁴⁹.

We generated the “average footprint” for the sequence harbouring this mutation and compared it to that of the reference sequence (Figure 64). The “average footprint” of the variant sequence corresponding to a GATA1 site shows a clear footprint compared to the reference profile, creating a high “damage score” with a negative value. Our approach has obviously predicted the formation of a footprint and highlighted a potentially interesting SNP to pursue, a SNP which was demonstrated to lead to the formation of an alternative promoter and affect α -globin expression.

In this section, we illustrated with two examples, one purely *in silico* and one previously functionally validated, how our “average footprints” “damage score” can be used to predict formation of new binding sites, which can be as detrimental to a regulatory landscape as a disruptive SNP

A)



B)

5'-TTATTA-3'

3'-TTATCA-5'



Figure 64. The Vanuatu mutation leads to formation of a new footprint in the alpha globin locus. A) Average profiles of reference and variant sequences. The mutation introduces a change in sequence that leads to formation of a footprint. B) The Vanuatu mutation creates a GATA1 site on the reverse strand, shown to impact on α -globin expression.

5.04 Conclusion

In summary, the implementation of “damage score” analysis allows us to assess and predict the effect of sequence variation on formation or disruption of footprints in a region of interest. High “damage scores” with positive or negative values are observed and indicate alteration or formation of TF binding sites respectively.

We characterise one of the variants with the highest damage score and predict a damaging effect on NF-E2 binding at R2/HS-40. With the aim of functionally validate this observation, a patient (RE) with unexplained anaemia and this precise variant is analysed. Preliminary experiments on the patient sample shows no obvious defect in erythropoiesis or in the formation of HS regions in the α -globin cluster. In order to confirm the effect of the variant on the binding site, we need to analyse the DNase digital footprints.

Finally, the R2/HS-40 base-pair change with the lowest negative damage score shows how this approach can be used to predict the formation of a new binding site. We validate this by testing the effect of the Vanuatu mutation and show that the variant creates an “average footprint”.

We illustrate how DNase-Seq data analysed with our novel approach (Chapter 4) of “average footprints” can be used to prioritise variants for functional analysis. By assessing the effect alterations of the sequence have on the potential of a sequence to be bound by TFs, we predict if a variant is damaging or creating a binding site. We believe this type of approach, that integrates tissue specific data with high resolution at regulatory regions, could be applied for prioritising functional validation of SNPs associated with common disease.

IV. Discussion

Summary

The scope of this study was to identify changes in the regulatory regions bound by protein, in different dynamic contexts such as development, and differentiation. The integrity of regulatory sequences and how they are decoded by TFs is increasingly important as alterations of the sequences underlying *cis*-elements is thought to increase susceptibility to common disease.

We used open chromatin data to identify all regions bound by TFs in a cell type. We analysed the HS regions (DNase-Seq and ATAC-Seq) at low and high resolution to assess the changes in sequence usage, employing different footprinting approaches (digital footprinting and the novel average footprinting).

At the α -globin locus, we show that quantitation of hypersensitivity can highlight significant changes in the levels of open-chromatin enrichment at different stages of development and differentiation. These changes are correlated with the activity of the gene promoters, even though direct binding at the α -globin enhancers might not change in certain contexts (i.e. development).

At a genome-wide level, the HS regions reflect a changing enhancer pool between developmental stages, and differentiating cells. We find the regulatory landscape in development to reflect different usage of the TF binding sites, based on “average footprints”.

Having a set of constrained sequences identified as bound by “average footprints”, we test *in silico* if alterations of these sequence will impact on TF binding. By assessing potential damage of binding sites, we highlight sequences where variation can impact on protein-binding, and affect gene expression. This approach is of utmost importance for the field of GWAS, where discriminating the myriad of associated non-functional from functional variants is a major challenge. Using the damage score of average footprints we can assess and prioritise associated variants for functional validation, based on their location in a bound site.

Finally, we present the case of a patient who has a variant in an enhancer of α -globin. Our predictive approach ranked this variant as potentially damaging. We present preliminary work to assess if, and how this variant impacts on α -globin expression, as an example of how SNPs in noncoding regions can affect gene expression.

Isolating regulatory regions- Open chromatin assays

Identifying the *cis*-regulatory landscape and the full complement of TFs they bind is crucial for understanding gene regulation in a cell of interest. One approach that allows for a comprehensive interrogation of sequence occupancy, independent of the identity of TFs, is to isolate open chromatin regions. DNase-Seq, based on the classical method of DNaseI assay, is the gold-standard in the field. Being a tedious protocol that requires high input material, DNase-Seq was recently overshadowed by a simpler new method, ATAC-Seq, which uses a hyperactive version of the transposase Tn5 to isolate the HS regions, and requires low input material.

We applied the two methods, in the same cell type, definitive erythroid cells and compared the output.

ATAC-Seq can reliably and more efficiently isolate open chromatin regions resulting in an overall higher HS signal than DNase-Seq at the same region. The ATAC-Seq libraries have more DNA complexity and result in better signal to noise ratio. This is not surprising considering that the transposase performs *in vivo* in a single reaction three of the most challenging steps of the DNase-Seq protocol: cutting of open chromatin fragments, blunting of ends, and ligation of adaptors. These steps all take place within the nuclei, followed by DNA isolation and amplification, eliminating in the process the stages where the biggest losses in a library preparation occur. The entire protocol is significantly more rapid than DNase-Seq, and most importantly, it can be applied to low cell numbers. We tested ~60000 cells which is 1000 fold less material than what was used for DNase-Seq. However, ATAC-Seq footprints (both digital and average) do not resolve binding sites to the same resolution as DNase-

Seq, and show sequence bias. There are several factors that may affect the resolution of ATAC-Seq digital footprints as compared to DNase-Seq ones. First we address potential biological aspects that might cause the observed bias in the data, then we follow with the technical aspects.

(a) Sources of biological bias in ATAC-Seq footprinting data

(i) Complexity of input material

The input material in ATAC-Seq is approximately 100-fold less than that used for DNase-Seq. This will result in lower complexity of the DNA within the HSs that are used for footprinting.

When analysing digital footprints, higher sequencing depth does not resolve the resolution (Results Chapter 1 -Footprinting), which emphasises the issue of library complexity.

When deriving “average footprints”, the ATAC library complexity aspect should be partially solved. The “average footprints” approach relies on a pile up of digital footprints, thus adding complexity to a single site based on its occurrence in the genome. This observation is confirmed to a certain extent, as we can detect the pattern of an “average footprints” in ATAC-Seq for erythroid specific motifs such as E-Box or GATA (Chapter 4). However, the resolution (width) and depth is not as pronounced as the DNase-Seq complements, hinting to a bias introduced by the enzyme cutting mechanism.

(ii) Mechanism of cutting open chromatin

The mechanism of cutting through hypersensitive sites in chromatin differs between the DNaseI and Tn5. DNaseI inserts into the minor groove of the DNA and generates single-stranded nicks through hydrolysis¹⁵⁶, whereas the modified Tn5 forms a synaptic complex with the DNA before cleavage¹⁵⁷. The formation of the synaptic complex by the transposase is not clearly understood but involves twisting the DNA which may result in steric hindrance and could displace the bound TFs.

This displacement of TFs may be the reason why the high resolution data appears noisy in digital footprints (Chapter 1 & 3). The fact that CTCF, a highly conserved TF which binds to DNA through 11-zinc fingers¹⁵⁸, still forms a footprint, is probably an indication that strength of binding to DNA is a factor influencing the ATAC-Seq digital footprints output.

When examining “average footprints” profiles, the bias introduced by the Tn5 cutting mechanism is highlighted by 2 aspects; an A/T sequence bias and wider profiles of binding observed in ATAC-Seq average footprints. The sequence preference has been reported previously when studying wild type transposase¹⁵⁹, and was suggested to be caused by the ease of distortion of the A/T regions. Being able to bend the target DNA is essential for the Tn5 mechanism as mentioned above^{157,159}. The wider footprint pattern might be an indication of protein displacement in the cutting process, thus losing the resolution of the footprint on the DNA. The fact that this is visible in the average footprints suggests that this displacement must happen consistently, at a genome-wide level.

(iii) Technical bias in ATAC-Seq footprinting data

In terms of technical causes that could affect the footprints, there is one outstanding aspect of the transposition that is yet to be modelled in computational analyses and applies to both digital and average footprints. When cutting the target DNA, the transposase generates a 9bp duplication region at the end of a DNA fragment. The authors of ATAC-Seq suggested this 9bp duplication be dealt with by subtracting 4 and 5 bp from each end of mapped fragments⁴⁴. This would remove the 9bp by shortening the fragment from both ends. While this minor technical detail has no impact on mapping hypersensitive sites with 250-300 bp resolution, it can have a significant impact when processing footprints at a base pair resolution resulting in shifting fragments and creating artefacts.

We are currently testing different shifting of the 9bp (combinations of 0-9, 1-8, 2-7 etc) or ignoring the 9bp duplication. The latter has resulted in promising preliminary results but still requires further investigation.

In conclusion, ATAC-Seq is a recent development, and as such, further characterisation of the mechanism the mutated Tn5 processes chromatin (all the characterisations available test Tn5 on genomic DNA) is needed to allow for proper computational analysis of the data output. Currently, to generate footprints, ATAC-Seq data is processed following the same pipeline used for DNase-Seq. Considering the fundamental differences in the enzymes' characteristics, it is unsurprising that the output from ATAC-Seq is generally inadequate.

ATAC-Seq is a recent development, and as such, further characterisation of the mechanism of the mutated Tn5, and the way it processes chromatin (all the characterisations available test Tn5 on genomic DNA) are needed to allow for proper modelling of the data. Currently, we are processing ATAC-Seq data in the same manner as DNase-Seq data to generate footprints, which considering the mechanistic differences in the methods might not be the most appropriate. With the current understanding and processing of footprints we cannot generate the same quality footprints from ATAC-Seq data, exception being proteins strongly bound to DNA such as CTCF.

(b) DNase-Seq footprints remain the ‘gold standard’

One might ask which of the two assays should be trusted. It is of course known that the DNaseI also has biases in the way it cleaves the DNA. These biases have previously been described: DNaseI prefers to cut at the minor groove and is susceptible to some extent to methylation¹⁶⁰. It has also been shown that these can be controlled for, by using the genomic DNA control.

Many arguments support DNaseI footprinting.

(1) Data obtained from this analysis recapitulates findings reported by other methods over the years (in vitro footprinting and DMS footprinting^{42,43}. (2) Moreover, DNaseI footprinting data agrees with the biology of the cell analysed. Our DNaseI “average footprints” approach highlights motifs such as E-box and GATA as strongest and most significant in the erythroid populations we studied, motifs known to be crucial in erythroid biology as reported by us and others following various methods^{61,153}.

(3) The DNaseI assay is well-established and characterised extensively in the last 30 years which allows a better modelling of the output and adaptation for genomics.

There is a possibility, as with other techniques, that DNaseI-Seq might not identify all the regions bound by protein (by missing highly dynamic proteins, or proteins easily displaced) which consequently increases the method's false negative rate (discussed below). However, from our experience, and compared to the alternative (ATAC footprints), DNaseI footprints remain the method of choice in identification of protein bound regions at high resolution. Combining this data with other types of data and methods is key in elucidating the biological question and compensating for the caveats that individual approaches present.

α -globin cluster at different developmental stages- Primitive and Definitive cells

(a) HS regions are similar in the two cell types

We compared HS regions in primitive and definitive erythroid cells, two distinct cell types with shared features: they are both red blood cells that synthesise haemoglobin for oxygen transport. One of the main characteristics of these two cell types is the expression of different globin genes.

We show, for the first time, the pattern of accessible chromatin in primitive erythroid cells at the α -globin locus, and genome-wide. The HS region of the ζ -globin promoter has one of the strongest enrichments of open chromatin genome-wide and is indicative of its transcriptional activity. Interestingly, the HS enrichment at the ζ -globin promoter is quantitatively stronger than the HS enrichment at the α -globin promoters. The observation highlights the functional collinearity at the α -cluster, as the arrangement of the promoters in the locus correspond to their HS level at the corresponding stage of development. The importance of the linear arrangement of elements and the 'promiscuity' of the enhancers in activating accessible HS regions upstream of the α -globin promoter has been demonstrated in the case of the Vanuatu mutation¹⁴⁹. This mutation creates a GATA1 binding site that leads to the formation of a promoter-like element upstream of the α -globin promoter, diverting the enhancers away from their target promoter and resulting in reduction of α -globin expression to pathological levels (causing HbH disease).

(b) Which enhancers are involved in ζ -globin regulation ?

Analysis of open chromatin at low resolution reveals that the five α -globin enhancers are hypersensitive at comparable levels in both cell types. This halts our attempt to predict enhancers' input in the regulation of both ζ -globin and α -globin expression in primitive cells based on HS profiling. In definitive erythropoiesis, we know that R1 and R2 are the main regulatory elements and R4 the minor regulatory element (Introduction, α -globin locus in the mouse). The role of R3 and Rm are not fully understood and we postulated they might be needed at a different developmental or differentiation stage. We will resort to genetics and test the contribution of all enhancers to transcription of ζ -globin gene by analysing the available knockout models in primitive erythroid cells.

We also tried analysis of HS at high resolution to determine if different enhancers' contribution to transcription of ζ -globin gene can be inferred from the sequence occupancy by TF using DNase digital footprints. Interestingly, the direct TF binding at the enhancers and the α -globin promoters is strikingly similar in both cell types at both stages of development.

Based on this data, we hypothesise that differences in regulation of the embryonic and adult gene in the cluster are mediated at the level of protein-protein interactions or 3D chromatin interactions, rather than direct TF binding. A comparison between primitive and definitive cells in the interaction landscape at the cluster has been done by Dr Duantida Songdej for her thesis study (Songdej et al., unpublished). The compartment of interactions is highly comparable suggesting a model in which the interactions are setup in a similar fashion in both tissues. Slight variations in the

intensity of interactions over the promoter of choice for a particular cell type could be caused by yet undetermined cues at different HS regions perhaps at the ζ -globin promoter. This type of fine-tuning within the well-defined compartment of interactions has been proposed at the β -globin in the form of 'active chromatin hub' as a model to explain the switching phenomenon¹⁶¹.

(c) Does the cue reside in the ζ -globin promoter?

In order to characterise the HS region of the ζ -globin promoter at high resolution, we applied the novel approach, the "average footprints" (Results Chapter 4) to scan the region of the promoter and identify the motifs that present a footprint in primitive cells. We then identified the most used sequences and generated the predictions of TF bound to those sequences based on available databases (Appendix).

Though we can predict the potential protein families potentially bound to the motifs discovered at the "average footprints", to identify the specific TF, this approach must be combined with other genomic tools. The validation of the predicted TFs requires RNA expression profiles in primitive cells, and more importantly, binding of proteins profile by ChIP methods. Usually the limitation of using ChIP experiment for exploring binding at a new site is the need for *a priori* knowledge of what factor(s) to test. We solve that problem by shortlisting candidates based on footprints. It would be of course interesting to compare binding at ζ -globin with binding at the ε -globin and β H1-globin in the β -globin cluster, to identify similarities and differences in the regulation that might provide insights into the phenomenon of "switching".

α -globin cluster at different stages of erythropoiesis - early progenitors and differentiated cells

As erythroid differentiation progresses, adult α -globin expression increases. We observe that all the HSs are present in the cluster at early and late stages of differentiation, though their levels of accessibility, visualised and quantitatively assessed from peak height, differ. There is an overall reduction in hypersensitivity in the early progenitors, with significant quantitative reduction at R2, R3 and the adult α -globin promoters.

The novelty in our observations resides in the fact that we quantified the level of open chromatin enrichment and found it reflects the transcriptional activity of the promoters, and potentially the sequential binding of the enhancers to the promoters seen during differentiation (Anguita et al).

The transcriptional activity of the α -globin genes is most probably the result of cooperative TF binding at the enhancers following the opening of chromatin. We have shown (Hay, under review) that the major regulatory elements, which have a functional contribution to α -globin transcription, are R1 and R2. Both enhancers need to be setup for full transcriptional activation, as seen in differentiated definitive cells in WT and confirmed in the double KO model (Hay, under review). In the progenitor population, R2 is significantly less hypersensitive, likely suggesting that not all the necessary TFs are bound. This observation would be best confirmed by analysing the footprints in R2, which we were unable to finalise due to the limitations in

footprinting using ATAC-Seq data. Thus, this aspect would need to be addressed using DNase-Seq digital footprints, for which input material is still a limitation.

The role of the R4 and Rm enhancers in erythroid cells is unknown- they may act as 'shadow' enhancers, ensuring robust expression (Barolo 2012), or might be needed at a different stage in development.

(a) Enhancer priming in the α -globin cluster

The emerging picture from the present study and others^{53,81} is one of enhancer priming at the α -globin cluster in the early differentiation stage. Thus, the presence of HS regions is not only indicative of an element involved in active transcription, but may occur as a result of related TFs binding to 'label' the enhancers that will be necessary at a different stage (i.e. differentiated erythroid cells, or potential remnants from hematopoietic progenitors). This would prevent nucleosome repositioning and reduce the rate limiting step of chromatin re-modelling¹⁶.

Such enhancer priming is observed in ES cells, where members of the forkhead proteins bind the albumin enhancer sequentially: early binding of FOXD3 prevents DNA methylation and ensures FOXA1, required at later stages of development, can bind¹⁶².

Similarly, some of the α -globin enhancers could be hypersensitive as remnant of their activity in previous developmental stages. For example, we observed an interesting fact in myeloid progenitors (416B cells): Rm and R2 are hypersensitive, with Rm being significantly more enriched in open chromatin, while no other

erythroid specific HS is present (ENCODE data). Thus, two of the erythroid specific enhancers, one major and one with no observed involvement in α -globin transcription, are already primed at an earlier hematopoietic stage. Similar to the FOX TFs described above, related TFs are involved in binding these two enhancers in a stage specific manner: GATA2 at the hematopoietic progenitor stage, replaced by GATA1 at the committed, early erythroid stage⁵³.

These examples correspond to so-called ‘lineage-priming’¹⁶³ as enhancers are prepared for subsequent developmental stages, but there is no reason why this could not be the case in differentiation, where arguably the transitions are even more rapid. Thus, R2 and R3 could be primed in the early progenitors, and only become involved in α -globin transcription in differentiated cells.

Based on the finding that the α -globin enhancers are already ‘primed’ in hematopoietic progenitors, it is not unreasonable to propose that studying the enhancers setup in the transition from closed chromatin, to fully transcriptionally active α -globin locus has to start from earlier hematopoietic stages and not be constrained to a committed erythroid population. It appears more likely, In the light of the commitment/differentiation process being a continuum, in which even a cell-cycle might make a difference¹⁶⁴, that the locus is sequentially setup and bound from an earlier developmental stage.

(b) Signature to distinguish primed enhancers from functional enhancers

Correlated with the enrichment of the HS regions at the α -globin enhancers are some of the chromatin marks at the two stages, with H3K27Ac having a strikingly stronger pattern in differentiated cells. This is in agreement with the observation that K27Ac can distinguish between active and inactive enhancers^{101,119,121}. However, the acetylation domain is quite broad and there is no distinction in the level of acetylation enrichment between the functionally active enhancers (probably R1) and potentially primed enhancers (R2, R3).

In terms of the other histone marks, we observe an increase in H3K4me1 at the two enhancers that become significantly more hypersensitive in differentiated cells, R2 and R3, and of course in the levels of H3K4me3 at the promoters.

We observe a similar pattern of histone modifications at two other loci that become active in differentiated cells: β -globin and Mitoferrin1.

Thus, a refined signature for active enhancers could be derived that includes quantitative measurement of HS regions enrichment and H3K4me1 levels at individual enhancers, in addition to K27Ac. This would help distinguish active elements in clusters of enhancers as opposed to primed or not yet functional at a specific stage.

In order to validate these observations we would need triplicates for the chromatin marks in the two population for quantitative analysis, and testing of this signature at loci where we understand the dynamics of gene activation, similar to the analysis we performed at the α -globin locus.

(c) Development and differentiation of erythroid cells- Genome-wide analysis

The two dynamic contexts, development and differentiation, in which we analysed the changes in HS regions share a common pattern: we identified approximately 10% cell specific HS regions. The differential open-chromatin regions correspond mainly to enhancers, but also promoters. Majority of these regions (~80%) occur in different genomic regions. These observations reinforce the fact that differentiation and development are driven by TF binding at distal elements.

Our observations on the dynamics of open chromatin are consistent with the findings of a broader comparison of DNase-Seq hypersensitivity in HSC cells and differentiated human cell lines³⁷. In this study, they identify a core set of HS regions which is common between ESC and developing cell types, and decreases during differentiation. Apart from the core set, loss and gain of cells specific HS regions is observed as the cells progress through development. This analysis was performed on much wider developmental time-scale, using ESC and differentiated cells, by combining the DNase HS regions from multiple cell lines. As such, the specifics of a particular cell type are not available. However, an important conclusion from this study, is the emphasis on the use of the dynamics of HS regions at distal elements, to assess developmental progression even though the TFs driving these changes were not analysed..

A more recent study addressed the dynamics of chromatin modifications at distal elements throughout hematopoiesis using ATAC-Seq⁹². Unlike the previously

described paper, this study investigates the development of one specific system, hematopoiesis. The loss and gain of enhancers is assessed, from HSC to differentiated cells in all the lineages. Interestingly, the erythropoietic lineage loses the largest proportion of enhancers from HSC. The factors bound at the erythropoietic HS regions are assessed only based on known motifs enrichment and, as expected, known erythroid TFs are identified (GATA1, KLF1, SOX6).

Though they highlight the dynamics of open-chromatin landscape, neither of the two studies assesses in a systematic manner the drivers of these changes- either the analysis is done at a global level, in multiple cell lines losing thus the tissue specific aspect, or, if in a specific cell line, it is performed in a biased manner, based on known motifs and no information about binding at those motifs.

In Chapter 4, we show how analysis of open-chromatin regions can be used to explore the binding patterns that drive cell specific differences. We use average footprints to incorporate information about bound sequences and identify the TFs involved in the dynamics of the non-coding genome, and in the regulation of gene expression. Our analysis of average footprints incorporates information on bound sequences, highlighting sequence usage, and ranks significantly bound motifs. More importantly, the approach is completely unbiased, and allows for identification of *de novo* bound regions and thus one can identify new TFs.

The usage of open-chromatin landscapes can provide insights into the activity of the non-coding genome and how regulatory regions are setup and coordinated in a specific cell type. We show that, apart from identifying the changes of HS regions, a

more informative approach that improves on resolution, is to use footprinting information.

Protein-DNA interactions- Factors involved in TF binding to regulatory regions

Regulatory sequence encode information for correct temporal and spatial development. Readers and interpreters of this information are TFs, contribute to the initial steps of transcriptional regulation by binding cell specific regulatory regions and recruiting additional factors and finally the transcriptional machinery.

Despite the recent progress in high-throughput technologies that allows to assess TF binding at large scale and in a fast manner, conflicting results from in vitro and in vivo assays in terms of protein binding affinities, motif recognition or dynamics make it difficult to untangle rules of TF binding at regulatory.

We show through this current study, that the use of high resolution DNA accessibility can clarify some of these aspects of TF binding to DNA and can be used for further understanding gene regulation

We will systematically go through steps the hypothetical TF binding to a regulatory region, an enhancer, and discuss each stage in the light of high resolution profiles, footprints, and how they can be complemented with other data for mechanistic insights.

(a) TFs bind to regions of accessible chromatin

The fact that TFs binding is occurring at accessible chromatin has been the basis of assays such as DNaseI digestion. Our study of binding at high resolution focuses on the aspect that TF footprints are found in HS region. However, if the chromatin is already open when TFs bind, or if they lead to accessibility of the DNA, is not fully established.

A special class of factors has been described as pioneer TFs that are able to bind nucleosomal DNA *in vitro*¹⁶⁵, and relatively condensed chromatin *in vivo*. These pioneer TFs are believed to have unique properties that allow them to modulate nucleosomal density and increase accessibility for other factors. However, it is becoming clear that pioneer factors are not binding to heterochromatic sites, but they are the first to bind to transient, decondensed sites that are not accessible to other TFs^{130,166,167}. Well-characterised pioneer factors include members of the Forkhead box proteins, such as FOXA shown to bookmark enhancers by reprogramming the surrounding chromatin during neuronal development^{130,168}.

Interestingly, when analysing the average footprints in differentially enriched regions in primitive and definitive cells, we find the Forkhead box protein family are a frequent prediction for the most used motifs. The top six motifs in primitive cells (Figure 50, Results Chapter 4) and motifs ranked 10th to 15th in definitive cells include members of this family.

In contrast to the pioneer factors, 'settler' TFs cannot open chromatin and are highly dependent on the accessibility of the DNA underlying their motif. Moreover, settler TFs can increase the chromatin accessibility upon binding¹⁶⁹ (Figure 65).

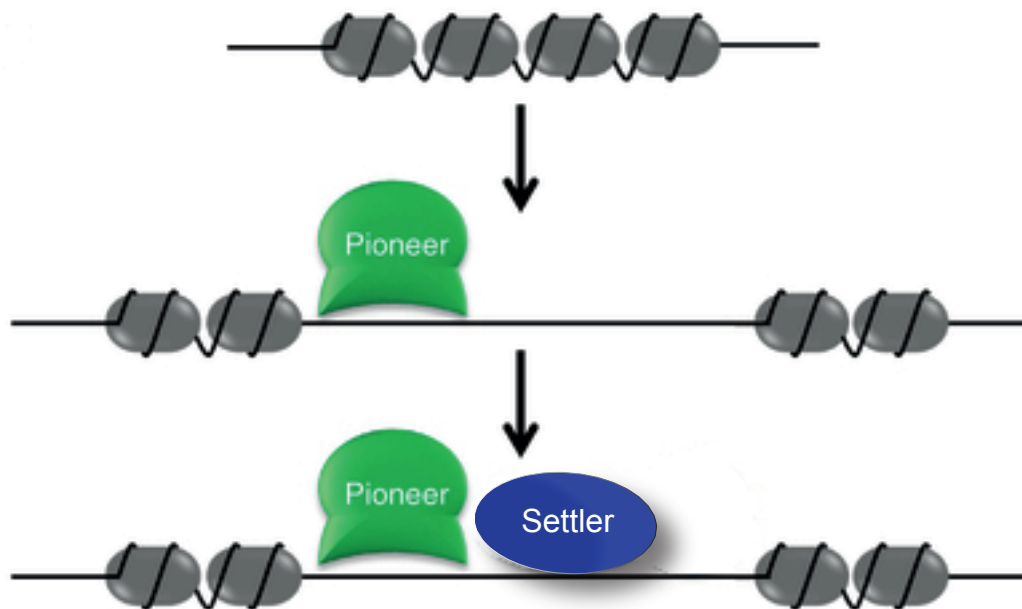


Figure 65. Illustration of hierarchy of binding of TFs at HS regions. Model of TF binding hierarchy. Pioneer TFs binding initially can open chromatin, allowing for settler TFs to populate the site. Figure adapted from (Hashimoto et al., 2014)

In erythropoiesis, it is not clear which factors play “pioneer” roles, though the order of events has been addressed in the GATA1-induced differentiation system¹⁷⁰. They show the chromatin state and accessibility remain largely unchanged and are not dependent on GATA1. Based on the observation that SCL/TAL1 precedes GATA1 binding at many sites in proliferating progenitors^{71,170}, it might be that TAL1 has pioneer function when directly binding to DNA.

The cell-specific HS regions we identified in development and differentiation consist both of HS sites unique to one condition, and also of sites that differ in the amount of enrichment. One can imagine a scenario where some of the sites are established by pioneer TFs and are thus enriched for such motifs, whereas other sites have already been made accessible by pioneer factors and are subsequently bound by settler TFs, who recruit further factors and thus increase the level of open chromatin enrichment. We could test this aspect by further dividing the differential regions (new sites vs sites that present quantitative changes) and by extracting the average footprints. The usage of motifs would highlight the families of TFs involved in establishing new HS regions (‘pioneer’ TF) and in increasing HS regions accessibility (‘settler’ TF).

(b) Recognition of TF binding sites within a HS region is dependent both on sequence and structure

Once in the open chromatin regions, TFs, depending on their type, recognise their binding sequence through an interplay of base-readout and shape readout. Base-readout refers to the recognition of the chemical signature of the bases, through the physical interaction of the TF aminoacids and accessible DNA residues ¹⁷¹. Some families of TFs, such as bHLH type, rely mainly on this mechanism of DNA recognition ¹⁷².

Shape-readout refers to the recognition of the DNA physical properties at the binding site such as bending or minor/major groove read-out. TF families relying on this mechanism include the high-mobility group (HMG) type ¹⁷¹.

To date, genomic studies have mainly considered the base-readout mechanism as the main criterion for identification of TF binding. One of the reasons is probably the difficulty in assessing shape-readout with the current resolution of genomic tools. However, the role of physical structure in DNA recognition is re-emerging as a modulator of specificity.

One of the interesting questions is if the same family of TFs can recognise DNA alternating these mechanisms depending on context. If it turns out that TFs can use both recognition mechanisms, it might explain why certain binding sites identified by ChIP-Seq have no known underlying motif. By applying either the digital or the “average footprints”, we can pull out sequences that contain both types of readout. The hurdle is to discriminate between them. A surmountable challenge is to identify the TFs bound at the identified site. Methods such as high-resolution ChIP: ChIP-exo

²⁶ or ChIP-nexus ²⁷ could be coupled with footprinting in order to identify the specific TF occupying a certain digital footprint.

Binding through recognition of shape of the regulatory DNA might also account for the observed functional conservation at *cis*-elements without sequence conservation. In vertebrates, while binding sites can vary between species, function and trans-acting factors are conserved ^{173,174}. This could be explained both by affinity of TFs to multiple sites from a base-readout perspective, and also by shape-readout modulated by binding partners. It is only by improving the resolution of the methods that study protein-DNA interactions that we can address such aspects.

(c) How do TFs form footprints within the HS region?

Identification of site-specific protein binding is based on the principle that TFs will prevent nucleolytic attack on the contacted bases, by stable binding and thus, sterically blocking the DNaseI ¹⁷⁵. This process results in differential protection of nucleotides and specific cleavage patterns ^{33,38,96}.

However, *in vivo* studies of protein binding have shown that the residence times of TFs on the DNA, is relatively short, for example 10s for the glucocorticoid receptor ¹⁷⁶ and proteins move rapidly on and off template ¹⁷⁷. How do then TFs form footprints?

The relation between dynamics of binding and digital footprints was investigated recently and it was shown that the length of the residence time is proportional to the depth of the digital footprint pattern ¹⁷⁸. This observation is rather intuitive, as one would expect that a TF with short residence time would not provide sufficient DNA

protection from the highly processive DNaseI. It also became apparent, that different TFs, depending on their role in gene regulation can vary in their binding behaviours¹⁷⁹: factors involved in signalling processes such as GR, NF- κ B, p53 and ER will have very fast binding kinetics¹⁸⁰⁻¹⁸², AP1, intermediate¹⁸³, whereas factors involved in stabilising interactions such as CTCF will have much slower binding kinetics¹⁸⁴. The digital footprints left by these factors range from weak or non-existent (rapid kinetics), to very strong (slow kinetics)¹⁷⁸.

In red blood cells, erythroid-specific TFs probably have intermediate binding kinetics, as they leave clear digital footprints as illustrated by binding of GATA1, NF-E2 (through AP1) and SCL/TAL1 at the α -globin locus in both primitive and definitive cells. It is reasonable to assume that in order to form protein complexes such as the pentameric complex⁷², these TFs would stabilise the interaction with the underlying regulatory region. In terms of stronger binding, we inferred the same stable interaction for CTCF as previously described in our erythroid data (Chapter 1, Figure 26), with digital footprints detectable not only in DNase-Seq but also in ATAC-Seq data, being one of the only proteins stable enough to protect the underlying DNA from the transposase cutting. Even though we can detect erythroid-specific TFs when assessing protein binding globally, we must keep in mind the limitation of footprinting in situations of highly dynamic binding where we are likely to miss TFs with very fast kinetics.

Missing some of the highly dynamic events by footprinting might also constitute an advantage considering the mechanism employed by TFs to bind DNA. In bacteria, it was shown that TFs sample the DNA numerous times, prior to reaching their target site, in a combination of sliding and diffusion called facilitated diffusion¹⁸⁵. In

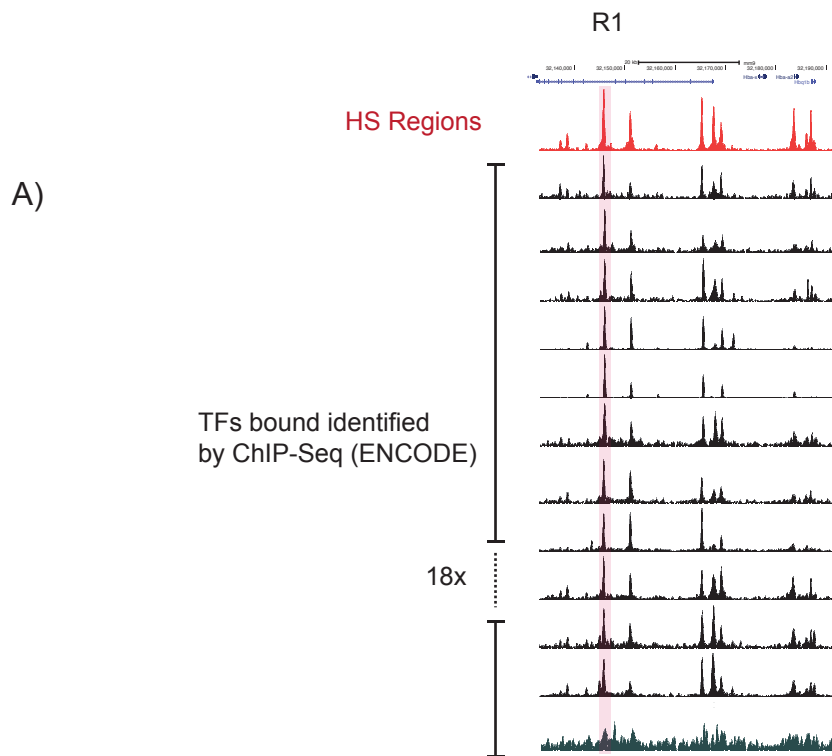
embryonic stem cells, using single molecule imaging of Sox2 and Oct4, it was demonstrated that sampling is essential for a TF to locate their binding site ¹⁸⁶. Ideally, we want to distinguish between sampling and highly dynamic, functional binding, which is not possible in a high-throughput manner.

In conclusion, TFs form footprints on the DNA by binding stably; the longer the residence time, the more likely to detect footprints. Highly dynamic binding as well as TF sampling have lower chance of being detected with current footprinting methods.

(i) Footprinting data overcomes some confounding effects of cell populations in genomic analysis

Finally, we discuss one confounding factor in current genomic analysis, the use of cell populations, and how the footprinting approach clarifies some of the data interpretation. Cell population studies result in an averaging effect; it can obscure dynamic and or sequential aspects of protein binding and creates a false image of abundance and co-binding of TFs. It has become evident that sometimes, the number of TFs bound at one HS region as identified by ChIP-Seq is difficult to reconcile with the available binding sites ³⁵, or the actual physical space available for binding. We tested this aspect by collaborating with a group (Narcis Fernandez-Fuentes, Aberystwyth University) who computationally models protein-DNA interactions. We modelled the binding at R1 (enhancer of the α -globin locus) of 18 TFs identified by ENCODE in an erythroid cell line (MEL) based on ChIP-Seq studies. Taking into consideration known binding sites, we could narrow the direct protein binding to 11 TFs. Initial analysis was performed and a theoretical model was built with minimal constraints: assuming the TFs would bind naked DNA in B-form,

and that there is no space needed in between the factors. Only using these parameters, it was possible to fit the 11 presumed direct binding proteins, out of the total 18 proteins that give ChIP signals at this enhancer (Figure 66). Based on the surface and volume calculations, it was concluded that the DNA surface of R1, if in chromatin form, should be 2-3 times bigger in order to accommodate all the mentioned factors without any steric or functional constraint.



B)

Modelling of TF binding at R1

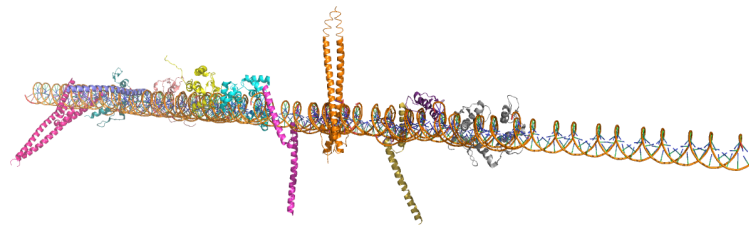


Figure 66. Binding of TFs as identified by ChIP-Seq at known enhancers (R1) and possible population effects. A) The R1 enhancer was found to be bound by 18 TFs in murine erythroleukemia cells. B) Theoretical modelling of those factors could only place 11 direct binders at R1, who cannot all occupy the site without steric hindrance. From left to right: pink-cMYC-MAX, blue-MAZ, purple-JUND, orange-GATA1, yellow-ETS1, turquoise- cMYB, pink- cMYC, orange- MAFK, yellow-MXI1, purple-ZNF-384, grey-ZK-SCAN1.

There are a few possible explanations for these observations: (1) differential binding in a heterogeneous population- different TFs are bound in different cell fractions and averaged out; the input of heterogeneous population affects all high-throughput genomic assay (including DNase-Seq, ATAC-Seq), and as such, is a confounding factor that cannot be addressed through same methodology, but rather with approaches such as imaging or single-cell assays, (2) indirect binding- some of the factors could be tethered to other factors, as suggested by the fact that not all observed TFs have known binding sites, and (3) non-functional binding- some factors might be in the process of sampling the DNA, or in the vicinity of a “hyperchipable” region¹⁸⁷ when crosslinking occurs.

We argue that combining ChIP-Seq data with DNase-Seq digital footprinting data can help tackle at least the second and third confounding factors described above. As a result, in the case of R1, we identify 7 sites that show high protection from DNaseI cleavage, including 2 GATA1 sites, 2 GATA-E-box sites and 3 uncharacterised sites, a more biologically feasible starting point compared to the 18 TFs reported previously in the attempt to unveil the key players underlying erythropoiesis in development and differentiation.

To summarise, if the interest is to understand the binding factors involved in gene regulation, some of the population heterogeneity effects can be excluded if using DNase footprinting data to complement ChIP-Seq data. Thus, by using information on the occupied regulatory sequence and the occupant TFs, one can control the variables affecting the process of interest, be that the direct binding or indirect binding.

(d) TF binding in the context of regulatory elements

TF binding sites are relatively short (6-12 bp), degenerate DNA sequences, essential upon recognition by TFs to enhancer activity. The arrangement of TF binding sites within a regulatory sequence is referred to as motif 'grammar' ^{16,188}, and specifies orientation, relative position, and spacing of TF binding sites on DNA. For example, if orientation of one TF binding site is reversed, the TF binding there cannot be recruited in a manner conducive of protein-protein interaction with binding partners; order of TF binding sites can affect the direct protein interactions in a similar manner, while incorrect spacing of motifs can lead to steric hindrance or positioning on opposite faces of the DNA helix. Motif grammar and the surrounding sequence affect how the TF binding site is recognised and occupied by TFs and binding partners (Figure 67).

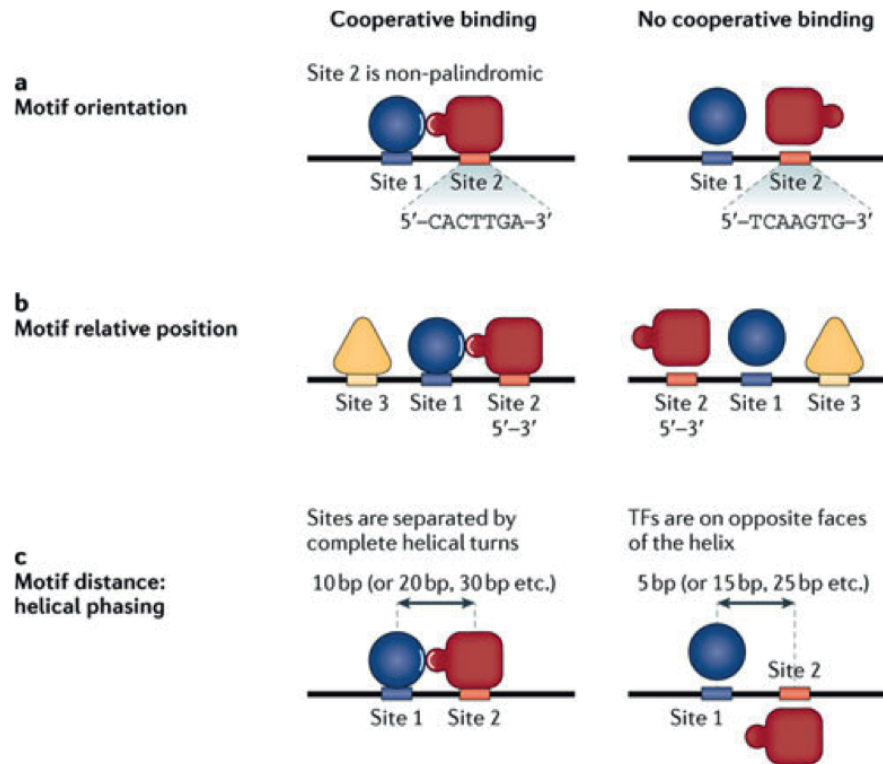


Figure 67. At regulatory sequences, TFs are subject to ‘motif grammar’. Diagram illustrating the factors that can affect ‘motif grammar’ within a regulatory sequence: a) Orientation of motifs can affect protein interactions and thus, cooperativity of the TFs. b) changing motifs position can also affect TFs interactions. c) Protein interactions are dependent on the distance and helical phasing of the motifs. Considering there are approximately 10 base pairs (bp) per helical turn of DNA, cooperative interactions are facilitated by spacing of ~10bp or multiples of this. Spacings such as ~5 bp, 15 bp or 25 bp will result in the TFs being positioned on opposite faces of the DNA helix, impairing their interaction. These examples illustrate cases where even though motif composition has not changed, cooperative binding can be impaired because of motif arrangement. This figure was adapted from (Spitz and Furlong, 2012).

High-resolution footprinting data provides a method to interrogate such aspects of gene regulation *in silico*. A combination of digital footprinting and average footprinting can be employed to identify the preferred motifs, orientation and positioning relative to other motifs at regions of interest. Moreover, binding preferences of specific TFs to variations of motifs can be interrogated and associated with enhancer activity.

For example, we notice that the footprints at R1 in the α -globin locus, though predominantly GATA binding sites, have different orientations and different depths, with the GATA-E-box motif in the reverse orientation forming stronger footprints. We could test the occurrence of the motifs in this configuration at other erythroid loci and correlate with the activity of the enhancers. This would involve combining the footprints, ChIP-Seq data and gene expression data. Ultimately, we could use computational analysis to learn the pattern of footprints from active enhancers, in order to predict the functional state of enhancers, genome-wide.

Though motif grammar is important for models of enhancer activity in the light of cooperative models (Introduction, Models of enhancer output), the TF motif positioning and spacing is rarely linked to enhancer activity. This could be due to the resolution of current genomic methods or the challenge in identifying active enhancers and their target gene to be dissected as model loci at this level.

However, regulatory sequence composition has been recently tackled by a few studies in *Drosophila*. One study changed the order and spacing of the motifs in the eye enhancer *spa* and managed to change the output from cone-cells to photoreceptors¹⁸⁹. Such experiments reinforce the fact that regulatory sequences

serve as platforms for regulatory inputs and can determine the nature of the expression output.

Another group designed synthetic enhancers with various combinations of known binding sites and generated stable transgenic *Drosophila* embryos in which the activity was assessed ¹⁸⁸. Their findings showed that motif composition can affect expression both in a binary manner (on/off) due to different spacing or positioning, and also in a quantitative manner by influencing robustness of expression. Finally, the most interesting implication of these finding on my thesis is the observation that the motif grammar disruption affected enhancers in a tissue specific manner: while in some tissues the enhancer could have a 'flexible' motif composition, in another tissue it could be subject to rigid constraints. This shows how the same sequence variation can impact on various tissues differently.

This leads us the final part of the discussion; the role of sequence variation at regulatory elements in disease.

Determining the role of variation in disease- challenges and solutions

Alterations of the linear sequence of regulatory regions affect essential interactions with the 'interpreters' (TFs) and have been now shown to increase susceptibility to common disease ^{11,190}. GWAS identify large numbers of SNPs associated with common disease. In order to determine the causal SNP, the associated SNPs must be functionally validated. Considering the number of variants are in the range of thousands ¹¹, this proves logistically prohibitive. Thus, there is a need to prioritise the candidates for functional analysis.

Prioritising candidate variants based on their potential effect on gene expression is challenging due to the complexity of the non-coding genome: (1) regulatory elements function in a cell specific manner ^{5,16,35}; (2) the elements interact with single or multiple target genes located at varying distances, upstream or downstream, or within gene bodies^{35,97,191}; (3) the function of regulatory elements depends on the surrounding chromatin state ^{53,57,192} and the configuration of TF binding sites ^{16,172,177,188,193}.

We will briefly discuss the challenges in determining the causal variant, from prioritising of candidate variants to validation, in a hypothetical workflow of validating a SNP, emphasising the part where this project has made a contribution.

(e) Identifying causative variants

1) Cell-type specific data

It is now evident that the impact of sequence variation effect on gene regulation should be assessed in its native context, in vivo, in a cognate cell type ^{11,35}. To this end, we explore sequence variation in the context of open-chromatin, where binding of TFs to regulatory regions is likely to be functional and potentially a target for perturbation in a cell specific manner. We showed how DNase-Seq and ATAC-Seq can be used to identify HS regions in a cell-type, genome-wide, reliably and with high specificity (Results-Chapter1).

Moreover, to facilitate prioritisation, we have developed an approach that can predict damage to a binding site. This method is based on DNase “average footprints” and can be used to score variants based on their predicted impact on protein binding. By using HS data and the “damage score” approach, we can first identify the variants in regulatory regions, and prioritise candidate variants for further studies.

2) Determining target genes

Assuming we have shortlisted potential causative SNPs found in HS regions based on “damage score”, we need to identify the target genes of these regulatory elements where the SNP resides. By investigating model loci such as the α -globin, in a cell type extensively studied like erythroid cells, this obstacle is overcome as the regulatory elements have long been identified and characterised ¹⁵³. However, a common disease, such as diabetes, presents a more complex picture, with multiple loci potentially contributing to the phenotype ¹⁹⁴. Thus, after identifying tissue-specific HS regions containing the prioritised SNP(s) associated with the phenotype, the

interactions of that regulatory region/SNP with the surrounding landscape can be explored with 3C methods²³. This can be achieved at high resolution and by interrogating multiple loci of interest simultaneously using the Capture-C method¹²². Once the interactions of the predicted candidates have been determined, we can proceed to validate the effect the variants have on the gene(s) they interact with in their native environment by genetic manipulations.

3) Validating the predicted effect by genetic manipulations

The final step in the process is the functional validation of the predicted variants, using options such as genetic engineering. Recent advancements in Crispr/Cas9 systems have made the process of manipulating sites extremely rapid¹⁹⁵. This allows for testing the effects of mutations *in vivo*, in the native chromatin environment, given the availability of model systems to analyse the appropriate cell population affected.

This would constitute a potential workflow for testing variants associated with a complex phenotype. The integration of open-chromatin data, in particular DNase-Seq, in the analysis, is a strong determinant of functional variation^{13,49} (Figure 68). DNase-Seq can highlight variants that fall within active regions of chromatin, addressing the two of the challenges of GWAS: (1) HS are tissue and cell type specific and thus can help filter the variants by highlighting the ones likely to be functional in the relevant cell type to the disease in question; (2) variants in linkage disequilibrium (LD) with the associated SNPs from GWAS can also be analysed in relation to their occurrence in active chromatin⁴⁹.

We believe this sort of analysis is very beneficial for the GWAS field, in particular when addressing multigenic diseases. In such cases, as compared to an erythroid phenotype, the associated variants and their linkage is more complex. By using the DNase-Seq step with the damage score, one can perform an initial screen, in silico, at a genome-wide level, which incorporates functional information. Thus, the result is a short-list of candidates, obtained in a resource efficient manner, prioritised based on their location in a protein-binding site. We argue our approach facilitates the downstream analysis, and thus increases the chances of identifying disease-causing variants.

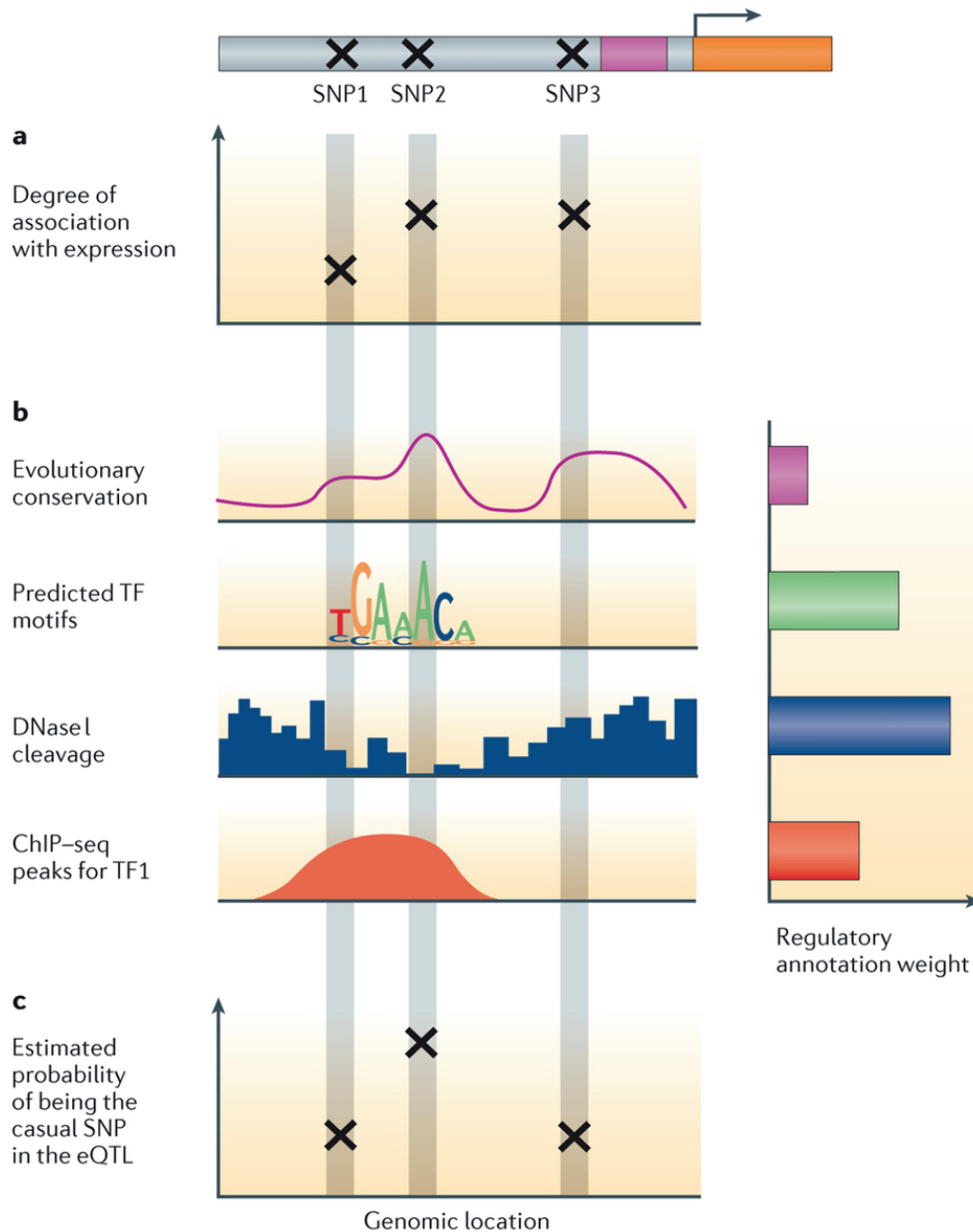


Figure 68. A possible scheme for the integration of GWAS information and regulatory annotation. a) Variants and their degree of association to the expression level the gene depicted. b) The same SNPs are also examined with respect to different regulatory annotations. Based on the annotations, probabilities for being the causal SNP can be assigned. The different regulatory features have weigh differently in predicting the causal SNP. c) Depiction of the probabilities of being the causal SNP by combining the information from parts a and b, such that even if two SNPs had a similar degree of association with gene expression from GWAS, one could receive a higher probability based on its regulatory annotations. Abbreviations: ChIP-seq, chromatin immunoprecipitation followed by high-throughput sequencing; TF, transcription factor; DNase I cleavage- DNase I footprints. Figure adapted from (Levo and Segal, 2014).

In Chapter 5, we illustrate our contribution to the field by using DNase-Seq data to assess the impact of sequence variation at the α -globin locus. Not only does DNase-Seq result in cell specific information, but it can also provide sequencing depth required for analysis of HS regions at high resolution. By adapting our novel “average footprints” approach to calculate “damage score” we demonstrate how DNase footprinting information can be used to predict the effect of a sequence variant.

A non-coding variant in the major regulatory element of α -globin

As proof of principle for the “damage score” approach, we have analysed *in silico* the effect of variants in R2/HS-40, the major regulatory element of α -globin. It is interesting, that based on the damage scores (Figure 59), two of the four highly conserved binding sites, the AP1 sites are predicted to be more affected by mutations. One of the patient referrals with unexplained anemia actually has a variant with a high damage score (rs177510) lies in one of these two AP1 sites.

We hypothesise that a variant in the conserved binding sites will affect TF binding and reduce α -globin expression and cause anaemia. We have strong reasons to believe that a variant in an enhancer can disrupt gene regulation, based on recent publications. A recent study showed how a recessive mutation in an enhancer element of *PTF1A* can cause isolated pancreatic agenesis. Functional annotation of the region with the associated mutations proved instrumental in understanding the causative effect ¹⁹⁶. In an erythroid context, common genetic variation associated with increased levels of haemoglobin was found in the non-coding regions

surrounding *BCL11A*. Functional annotation revealed a variant affecting a TF binding site, which reduces *BCL11A* expression and consequently increases HbF production

151 .

Thus, we proceeded to investigate the variants (rs177510) in R2/HS-40 *in silico* and *in vivo*, in the patient RE.

Variant lies in a binding site of NF-E2

The AP1 sites in R2 have been shown, by gel shift assay, to be bound by NF-E2⁴² at the extended AP1-motif: T/CGCT**G(A/G/C)TC**AT/C. NF-E2 is a basic-leucine zipper heterodimeric TF of the AP-1 superfamily. It is composed of the ubiquitous p18 and the hematopoietic-specific p45 subunits^{197,198}. Binding of both these subunits at R2/HS-40 was also confirmed *in vivo* by ChIP-Seq¹⁴⁶.

The role of the NF-E2 binding sites in α -globin transcription was investigated in the mouse, by replacing the two AP1 sites with LoxP sites. The mice were phenotypically normal, but had reduced allelic α -globin expression by 25%¹⁹⁹. This is a relatively mild phenotype. If the effect of the variant is similar to these changes in the mouse models, that might explain why the RE patient who has only one of the AP1 sites mutated has a mild anaemia. However, such conclusions from mouse models must be drawn with caution until functional validation. We have shown in the lab⁵² (Hay, unpublished data) that the sequence conservation of R2 does not translate into perfect functional conservation between mouse and human. For example, the deletion of the entire R2 in the mouse has a milder phenotype in terms

of α -globin expression than the deletion of R2/HS-40 in the human ⁵². This emphasises the need to validate the variant in human erythroid cells.

This is where applying the “damage score” for footprints can prove beneficial, as we can screen for a potential effect of the variant. The plots show that the variant sequence does not form a footprint when compared to the footprint seen in the unaffected or reference sequence, suggesting the new sequence does not have the potential to bind protein in erythroid tissue. Thus, the variant rs177510 has a high likelihood of affecting NF-E2 binding, affecting α -globin expression and causing the unexplained anaemia. We have proceeded with preliminary functional analysis of the variant in primary cells from patient RE.

Preliminary work and future plans

In order to infer from HS region formation if binding is affected at R2/HS-40, we performed a low-number open chromatin assay (ATAC-Seq) on cells from patient blood sample. The surprising observation was that the site appears as strongly hypersensitive as the control sample. This, however, does not imply that NF-E2 binding is not affected, but that disrupting the NF-E2 binding site does not disrupt the formation of the entire enhancer.

To address the state of NF-E2 binding to the sequence harbouring the variant (rs177510), a combination of digital footprints analysis and ChIP-Seq for NF-E2 will be performed on material derived from the patient RE. The fact that the patient is heterozygous for this variant (rs177510) offers an advantage from the analysis

perspective, as we can assess allelic skewing both in the footprints and in the binding.

To confirm the effect of the variant on transcriptional output, the α -globin expression as well as the ratio of α - to β -globin mRNA will be assessed in the patient in comparison to a normal individual,. Moreover, in order to link the effect of the variant to the reduced expression, we can assess allelic RNA expression, from the affected allele. Although the familial genetic would be the appropriate approach to assess the segregation of the mutation within the patient's family, the required samples were impossible to collect due to the personal situation of the patient.

Finally, to overcome the cell number limitations and dependency on patient samples which hindered the progress of this investigation so far, we are exploring the option of using genome engineering in immortalised human erythroid progenitors (HUDEP cells)²⁰⁰ to insert the point mutation and analyse its effect on α -globin expression *in vitro*.

V. Appendices and References

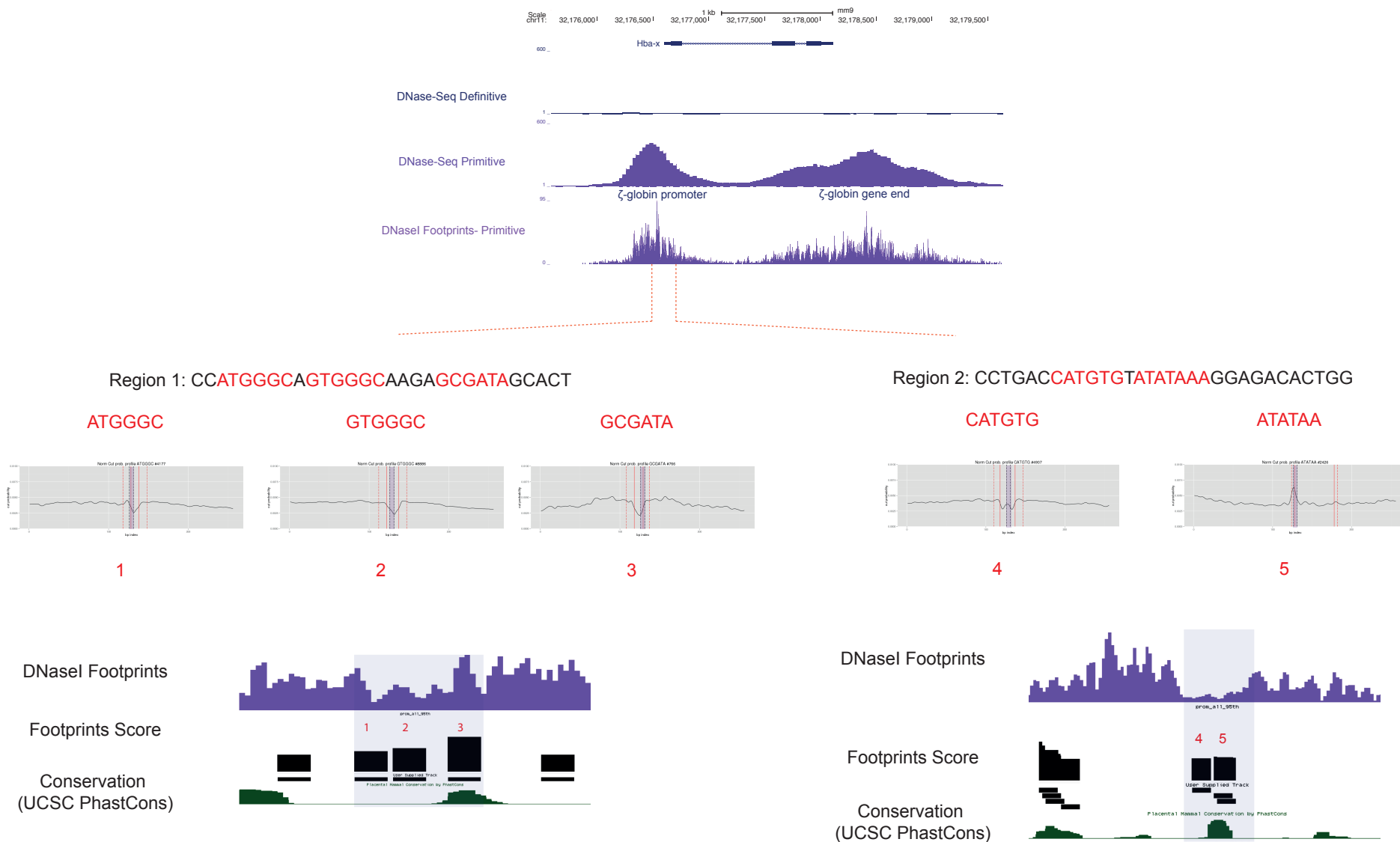


Figure 69. Examples of the strongest footprints at the ζ -globin promoter, identified using average footprints. Two sequences from the promoter containing three and two identified footprints, are shown together with the average footprint profiles as well as the single locus footprints. Scores of the footprints are represented as black bars, where the height of the bar is proportional with the footprints score. UCSC conservation is shown below the scores.

Table 4: Footprints at the ζ -globin promoter, sorted by average footprint strength with corresponding JASPAR predictions. Multiple entries represent multiple instances of the same motif identified within the promoter region.

Motifs	JASPAR Predictions
GCGATA	FOXL1=0.64; SPIB=0.62; NFIC=0.53; BRCA1=0.51; GATA3=0.51;
CTGATAAC	GATA3=0.63; SRY=0.6; MZF1_1-4=0.6; NFIC=0.6; FOXL1=0.6; SPIB=0.55; THAP1=0.51; BRCA1=0.51;
GCCCCC	MZF1_1-4=0.78; ZNF354C=0.73; THAP1=0.63; NFIC=0.56; MZF1_5-13=0.55; SP1=0.54; EBF1=0.54; INSM1=0.52;
GCGCAGCGCA	BRCA1=0.72; ZNF354C=0.59; FOXC1=0.59; SPIB=0.52; NRF1=0.51; MZF1_1-4=0.5;
TATCTC	FOXL1=0.69; SPIB=0.63; GATA3=0.63; SRY=0.51;
CCCCCT	MZF1_1-4=0.87; SPIB=0.64; MZF1_5-13=0.63; ZNF354C=0.62; THAP1=0.58; KLF5=0.58; FOXC1=0.57; EBF1=0.54; NFIC=0.53; SP1=0.53; INSM1=0.52; ZEB1=0.51; EHF=0.51;
GTGGGC	ZNF354C=0.9; NFIC=0.71; MZF1_1-4=0.66; THAP1=0.66; ZEB1=0.63; SREBF1=0.56; SREBF2=0.55; MAX=0.54; BRCA1=0.54; SP1=0.53; USF2=0.51;
GGCCCC	MZF1_1-4=0.67; NFIC=0.65; ZNF354C=0.63; THAP1=0.58; EBF1=0.52; KLF5=0.5; REL=0.5;
AATTCC	EHF=0.63; SPIB=0.6; NFIC=0.6; RELA=0.6; REL=0.59; NFKB1=0.55; MZF1_1-4=0.52; DUX4=0.51; ZNF354C=0.51; E2F6=0.51; FOXH1=0.51; E2F4=0.5; HINFP=0.5; NKX3-1=0.5;
ATGGGC	ZNF354C=0.73; NFIC=0.71; MZF1_1-4=0.65; THAP1=0.65; FOXL1=0.57; SREBF2=0.56; SPIB=0.56; SREBF1=0.56; MZF1_5-13=0.55; SP1=0.52; EBF1=0.52; CDX2=0.51; NRF1=0.51;
ATTGGC	NFIC=0.86; THAP1=0.63; SOX9=0.56; BRCA1=0.56; ZNF354C=0.56; FOXL1=0.53; FOXH1=0.52; SPIB=0.52; MZF1_1-4=0.52; DUX4=0.52; CDX2=0.51; SRY=0.5;
GCCCCCT	MZF1_1-4=0.77; SPIB=0.64; THAP1=0.62; KLF5=0.59; ZNF354C=0.58; MZF1_5-13=0.58; NFIC=0.56; SP1=0.54; EBF1=0.52; ZEB1=0.52; NRF1=0.51;
TTAGGT	FOXC1=0.64; SPIB=0.64; MZF1_1-4=0.6; NFIC=0.6; ZNF354C=0.59; NKX3-1=0.57; RORA_1=0.55; FOXL1=0.55; THAP1=0.54; DUX4=0.53; SREBF1=0.53; HINFP=0.51; GATA3=0.51;
CCAGCG	BRCA1=0.69; NFIC=0.65; ZNF354C=0.6; FOXC1=0.59; THAP1=0.59; MZF1_1-4=0.58; SPIB=0.54; NRF1=0.51;
GCCCAT	ZNF354C=0.73; NFIC=0.71; MZF1_1-4=0.65; THAP1=0.65; FOXL1=0.57; SREBF2=0.56; SPIB=0.56; SREBF1=0.56; MZF1_5-13=0.55; SP1=0.52; EBF1=0.52; CDX2=0.51; NRF1=0.51;
ATATAAA	FOXL1=0.68; GATA3=0.62; NKX3-1=0.59; FOXD1=0.56; SRY=0.55; FOXP2=0.54; CDX2=0.53; BRCA1=0.52; STAT1=0.51; FOXI1=0.5;
ATACAA	NFIC=0.7; NKX3-1=0.67; FOXL1=0.65; ZNF354C=0.65; SRY=0.65; SOX9=0.61; GATA3=0.61; BRCA1=0.59; FOXD1=0.59; FOXC1=0.58; CDX2=0.5;
GGCTCC	ZNF354C=0.65; NFIC=0.63; BRCA1=0.57; E2F6=0.52; MZF1_1-4=0.52; EBF1=0.52; E2F4=0.51;
TAGAGA	ZNF354C=0.65; SPIB=0.64; FOXL1=0.63; MZF1_1-4=0.61; NFIC=0.52; EBF1=0.52; GATA3=0.51; FOXC1=0.5;
CATGTG	MAX=0.62; NKX3-1=0.59; FOXC1=0.57; USF2=0.56; MZF1_1-4=0.56; BRCA1=0.55; USF1=0.55; ZEB1=0.55; SREBF1=0.53; ZNF354C=0.52; FOXL1=0.52; NRF1=0.52; MYC::MAX=0.51; EBF1=0.51; SPIB=0.51;
CATGTG	MAX=0.62; NKX3-1=0.59; FOXC1=0.57; USF2=0.56; MZF1_1-4=0.56; BRCA1=0.55; USF1=0.55; ZEB1=0.55; SREBF1=0.53; ZNF354C=0.52; FOXL1=0.52; NRF1=0.52; MYC::MAX=0.51; EBF1=0.51; SPIB=0.51;
CCCCTC	MZF1_1-4=0.79; ZNF354C=0.75; SPIB=0.66; MZF1_5-13=0.63; KLF5=0.6; THAP1=0.58; SP1=0.54; FOXC1=0.52; EHF=0.5;
CCCCTC	MZF1_1-4=0.79; ZNF354C=0.75; SPIB=0.66; MZF1_5-13=0.63; KLF5=0.6; THAP1=0.58; SP1=0.54; FOXC1=0.52; EHF=0.5;
CCCCTC	MZF1_1-4=0.79; ZNF354C=0.75; SPIB=0.66; MZF1_5-13=0.63; KLF5=0.6; THAP1=0.58; SP1=0.54; FOXC1=0.52; EHF=0.5;
TACAGT	FOXC1=0.73; SPIB=0.64; FOXL1=0.59; ZNF354C=0.58; SRY=0.56; NFIC=0.55; SOX9=0.53; THAP1=0.51;
ATTCCC	SPIB=0.69; MZF1_1-4=0.69; ZNF354C=0.65; NFIC=0.63; THAP1=0.61; RELA=0.59; REL=0.57; FOXL1=0.56; MZF1_5-13=0.54; NFKB1=0.54; STAT3=0.54; STAT1=0.53; E2F6=0.53; E2F4=0.53; FOXH1=0.52; EHF=0.51;

References

1. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Higgs, D. R., Vernimmen, D., Hughes, J. & Gibbons, R. Using Genomics to Study How Chromatin Influences Gene Expression. *Annu. Rev. Genom. Human Genet.* **8**, 299–325 (2007).
3. Lelli, K. M., Slattery, M. & Mann, R. S. Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.* **46**, 43–68 (2012).
4. Pepke, S., Wold, B. & Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* **6**, S22–S32 (2009).
5. Levine, M. Transcriptional Enhancers in Animal Review Development and Evolution. *Current Biology* **20**, R754–R763 (2010).
6. Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327–339 (2011).
7. Buecker, C. & Wysocka, J. Enhancers as information integration hubs in development: lessons from genomics. *Trends in Genetics* **28**, 276–284 (2012).
8. Cowper-Sal lari, R. *et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature Publishing Group* **44**, 1191–1198 (2012).
9. Sur, I., Tuupanen, S., Whittington, T., Aaltonen, L. A. & Taipale, J. Lessons from Functional Analysis of Genome-Wide Association Studies. *Cancer Res.* **73**, 4180–4184 (2013).
10. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
11. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
12. Sakabe, N. J., Savic, D. & Nobrega, M. A. Transcriptional enhancers in development and disease. *AA* (2012).
13. Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nature Publishing Group* **15**, 453–468 (2014).
14. Giorgetti, L. *et al.* Noncooperative Interactions between Transcription Factors and Clustered DNA Binding Sites Enable Graded Transcriptional Responses to Environmental Inputs. *Molecular Cell* **37**, 418–428 (2010).
15. Johnson, A. D., Meyer, B. J. & Ptashne, M. Interactions between DNA-bound repressors govern regulation by the lambda phage repressor. *PNAS* **76**, 5061–5065 (1979).
16. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nature Publishing Group* **13**, 613–626 (2012).
17. Thanos, D. & Maniatis, T. Virus induction of human IFN β gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995).
18. Panne, D., Maniatis, T. & Harrison, S. C. An Atomic Model of the Interferon- β Enhanceosome. *Cell* **129**, 1111–1123 (2007).
19. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry* **94**, 890–898 (2005).
20. Junion, G. *et al.* A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History. *Cell* **148**, 473–486 (2012).
21. Bulger, M. & Groudine, M. *Locus Control Regions (LCRs)*. 1–4 (John Wiley &

- Sons, Ltd, 2001). doi:10.1038/npg.els.0005034
22. Whyte, W. A. *et al.* Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153**, 307–319 (2013).
 23. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499–506 (2013).
 24. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
 25. Kim, T. H. & Ren, B. Genome-wide analysis of protein-DNA interactions. *Annu. Rev. Genom. Human Genet.* **7**, 81–102 (2006).
 26. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419 (2011).
 27. He, Q., Johnston, J. & Zeitlinger, J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology* **33**, 395–401 (2015).
 28. Higgs, D. R. *et al.* A major positive regulatory region located far upstream of the human alpha-globin gene locus. *Genes Dev.* **4**, 1588–1601 (1990).
 29. Follows, G. A., Dharni, P., Gottgens, B. & Bruce, A. W. Identifying gene regulatory elements by genomic microarray mapping of DNaseI hypersensitive sites. *Genome ...* (2006).
 30. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
 31. Song, L. & Crawford, G. E. DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols* **2010**, pdb.prot5384–pdb.prot5384 (2010).
 32. Sabo, P. J. *et al.* Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods* **3**, 511–518 (2006).
 33. Nepf, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
 34. Hosseini, M. *et al.* Causes and Consequences of Chromatin Variation between Inbred Mice. *PLoS Genet* **9**, e1003570 (2013).
 35. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
 36. Kowalczyk, M. S. *et al.* Intragenic enhancers act as alternative promoters. *Molecular Cell* **45**, 447–458 (2012).
 37. Stergachis, A. B. *et al.* Developmental Fate and Cellular Maturity Encoded in Human Regulatory DNA Landscapes. *Cell* **154**, 888–903 (2013).
 38. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).
 39. Galas, D. J. & Schmitz, A. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research* **5**, 3157–3170 (1978).
 40. Church, G. M. & Gilbert, W. Genomic sequencing. *PNAS* **81**, 1991–1995 (1984).
 41. Jarman, A. P. *et al.* Characterization of the Major Regulatory Element Upstream of the Human Alpha-Globin Gene-Cluster. *Molecular and Cellular Biology* **11**, 4679–4689 (1991).
 42. Strauss, E. C., Andrews, N. C. & Higgs, D. R. In vivo footprinting of the human alpha-globin locus upstream regulatory element by guanine and

- adenine ligation-mediated polymerase chain reaction. *Molecular and cellular ...* (1992). doi:10.1128/MCB.12.5.2135
43. Strauss, E. C. & Orkin, S. H. Guanine-adenine ligation-mediated PCR in vivo footprinting. *Methods* **11**, 164–170 (1997).
 44. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 1–8 (2013). doi:10.1038/nmeth.2688
 45. Gaffney, D. J., Veyrieras, J. B. & Degner, J. F. Dissecting the regulatory architecture of gene expression QTLs. *Genome ...* (2012).
 46. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7**, 862–872 (2006).
 47. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **368**, 20120362–20120362 (2013).
 48. Dickel, D. E., Visel, A. & Pennacchio, L. A. Functional anatomy of distant-acting mammalian enhancers. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, 20120359–20120359 (2013).
 49. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Research* **22**, 1748–1759 (2012).
 50. Flint, J. *et al.* Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Hum. Mol. Genet.* **10**, 371–382 (2001).
 51. Hughes, J. R. *et al.* Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *PNAS* **102**, 9830–9835 (2005).
 52. Anguita, E. *et al.* Deletion of the mouse alpha-globin regulatory element (HS - 26) has an unexpectedly mild phenotype. *Blood* **100**, 3450–3456 (2002).
 53. Anguita, E. *et al.* Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. *The EMBO Journal* **23**, 2841–2852 (2004).
 54. Weatherall, D. J. Haemoglobin and the inherited disorders of globin synthesis. *Postgraduate Haematology* (2005). doi:10.1002/9780470987056
 55. Viprakasit, V. *et al.* A novel deletion causing alpha thalassemia clarifies the importance of the major human alpha globin regulatory element. *Blood* **107**, 3811–3812 (2006).
 56. Phylipsen, M. *et al.* Thalassemia in Western Australia: 11 novel deletions characterized by Multiplex Ligation-dependent Probe Amplification. *Blood Cells Mol. Dis.* **44**, 146–151 (2010).
 57. Wallace, H. A. C. *et al.* Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. *Cell* **128**, 197–209 (2007).
 58. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369–375 (2012).
 59. PAULING, L., ITANO, H. A., SINGER, S. J. & WELLS, I. C. Sickle Cell Anemia, a Molecular Disease. *Science* **109**, 443–443 (1949).
 60. Williams, T. N. & Weatherall, D. J. World Distribution, Population Genetics, and Health Burden of the Hemoglobinopathies. *Cold Spring Harb Perspect Med* **2**, a011692–a011692 (2012).

61. Cantor, A. B. & Orkin, S. H. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene* (2002). doi:10.1038/sj/onc/1205326
62. Welch, J. J. *et al.* Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* **104**, 3136–3147 (2004).
63. Pevny, L. *et al.* Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. , *Published online: 17 January 1991*; | doi:10.1038/349257a0 **349**, 257–260 (1991).
64. Fujiwara, Y., Browne, C. P., Cunniff, K., Goff, S. C. & Orkin, S. H. Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *PNAS* **93**, 12355–12358 (1996).
65. Orkin, S. H. GATA-binding transcription factors in hematopoietic cells. *Blood* (1992).
66. Weiss, M. J. & Orkin, S. H. GATA transcription factors: key regulators of hematopoiesis. *Experimental Hematology* **23**, 99–107 (1995).
67. Siatecka, M. & Bieker, J. J. The multifunctional role of EKLF/KLF1 during erythropoiesis. *Blood* **118**, 2044–2054 (2011).
68. Tallack, M. R. *et al.* A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Research* **20**, 1052–1063 (2010).
69. Hall, M. A. *et al.* The critical regulator of embryonic hematopoiesis, SCL, is vital in the adult for megakaryopoiesis, erythropoiesis, and lineage choice in CFU-S12. *PNAS* **100**, 992–997 (2003).
70. Kassouf, M. T., Chagraoui, H., Vyas, P. & Porcher, C. Differential use of SCL/TAL-1 DNA-binding domain in developmental hematopoiesis. *Blood* **112**, 1056–1067 (2008).
71. Kassouf, M. T. *et al.* Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Research* **20**, 1064–1083 (2010).
72. Wadman, I. A. *et al.* The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *The EMBO Journal* **16**, 3145–3157 (1997).
73. Palis, J. Primitive and definitive erythropoiesis in mammals. *Frontiers in Physiology* **5**, 3 (2014).
74. PALIS, J., Robertson, S., Kennedy, M., Wall, C. & Keller, G. Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse. *Development* **126**, 5073–5084 (1999).
75. Kingsley, P. D., Malik, J., Fantauzzo, K. A. & Palis, J. Yolk sac-derived primitive erythroblasts enucleate during mammalian embryogenesis. *Blood* **104**, 19–25 (2004).
76. Gregory, C. J. & Eaves, A. C. Human marrow cells capable of erythropoietic differentiation in vitro: definition of three erythroid colony responses. *Blood* **49**, 855–864 (1977).
77. Gregory, C. J. & Eaves, A. C. Three stages of erythropoietic progenitor cell differentiation distinguished by a number of physical and biologic properties. *Blood* **51**, 527–537 (1978).
78. Higgs, D. R. & Wood, W. G. *Erythropoiesis. Postgraduate Haematology* 13–25 (Blackwell Publishing Ltd, 2005). doi:10.1002/9780470987056.ch2

79. Stamatoyannopoulos, G. Control of globin gene expression during development and erythroid differentiation. *Experimental Hematology* **33**, 259–271 (2005).
80. TomHon, C. *et al.* Evolution of a Fetal Expression Pattern via cisChanges near the γ Globin Gene. *J. Biol. Chem.* **272**, 14062–14066 (1997).
81. Vernimmen, D., De Gobbi, M., Stanley, J. A. S., Wood, W. G. & Higgs, D. R. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *The EMBO Journal* **26**, 2041–2051 (2007).
82. Higgs, D. R. & Wood, W. G. Long-range regulation of alpha globin gene expression during erythropoiesis. *Curr. Opin. Hematol.* **15**, 176–183 (2008).
83. Craddock, C. F. *et al.* Contrasting effects of alpha and beta globin regulatory elements on chromatin structure may be related to their different chromosomal environments. *The EMBO Journal* **14**, 1718–1726 (1995).
84. Bernet, A. *et al.* Targeted inactivation of the major positive regulatory element (HS-40) of the human alpha-globin gene locus. *Blood* **86**, 1202–1211 (1995).
85. Spivak, J. L., Toretti, D. & Dickerman, H. W. Effect of phenylhydrazine-induced hemolytic anemia on nuclear RNA polymerase activity of the mouse spleen. *Blood* **42**, 257–266 (1973).
86. Fibach, E., Manor, D., Oppenheim, A. & Rachmilewitz, E. A. Proliferation and maturation of human erythroid progenitors in liquid culture. *Blood* **73**, 100–103 (1989).
87. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research* **29**, e45 (2001).
88. Southern, E. Southern blotting. *Nat Protoc* **1**, 518–525 (2006).
89. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
90. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Research* **12**, 996–1006 (2002).
91. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728–1740 (2012).
92. Lara-Astiaso, D. *et al.* Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
93. Stark, R. & Brown, G. DiffBind: differential binding analysis of ChIP-Seq peak data. *In R package version* (2011).
94. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
95. Laing, E. & Smith, C. P. RankProdIt: A web-interactive Rank Products analysis tool. *BMC Res Notes* **3**, 221–3 (2010).
96. Boyle, A. P. *et al.* High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research* **21**, 456–464 (2011).
97. Maher, B. ENCODE: The human encyclopaedia. *Nature* **489**, 46–48 (2012).
98. Van Holde, K. E., Allen, J. R. & Tatchell, K. DNA-histone interactions in nucleosomes. *Biophysical ...* (1980). doi:10.1016/j.bpj.2011.07.057
99. Staynov, D. Z. & Proykova, Y. G. Quantitative analysis of DNase I digestion patterns of oligo- and polynucleosomes. *J. Mol. Biol.* **279**, 59–71 (1998).
100. Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* **8**, 1765–1786 (2013).
101. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect

- global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
102. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research* **21**, 447–455 (2011).
 103. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Research* **41**, e201–e201 (2013).
 104. Barozzi, I., Bora, P. & Morelli, M. J. Comparative evaluation of DNase-seq footprint identification strategies. *Front. Genet.* (2014). doi:10.3389/fgene.2014.00278/abstract
 105. Sankaran, V. G. *et al.* Developmental and species-divergent globin switching are driven by BCL11A. *Nature* **460**, 1093–1097 (2009).
 106. Bauer, D. E., Kamran, S. C. & Orkin, S. H. Reawakening fetal hemoglobin: prospects for new therapies for the β -globin disorders. *Blood* **120**, 2945–2953 (2012).
 107. Sankaran, V. G. & Weiss, M. J. Anemia: progress in molecular mechanisms and therapies. *Nature Publishing Group* **21**, 221–230 (2015).
 108. Huang, B. L. *et al.* Derepression of human embryonic zeta-globin promoter by a locus-control region sequence. *PNAS* **95**, 14669–14674 (1998).
 109. Yi, Z. *et al.* Sox6 Directly Silences Epsilon Globin Expression in Definitive Erythropoiesis. *PLoS Genet* **2**, e14 (2006).
 110. Kingsley, P. D. 'Maturation' globin switching in primary primitive erythroid cells. *Blood* **107**, 1665–1672 (2006).
 111. Yu, C. Y., Chen, J., Lin, L. I., Tam, M. & Shen, C. K. Cell type-specific protein-DNA interactions in the human zeta-globin upstream promoter region: displacement of Sp1 by the erythroid cell-specific factor NF-E1. *Molecular and Cellular Biology* **10**, 282–294 (1990).
 112. Sabath, D. E., Koehler, K. M., Yang, W. Q. & Phan, V. DNA-Protein Interactions in the Proximal ζ -Globin Promoter: Identification of Novel CCACCC-and CCAAT-binding Proteins. *Blood Cells* **24**, 183–198 (1998).
 113. Levy-Wilson, B. DNaseI hypersensitive sites at the 3' end of the human apolipoprotein B gene. *Biochemical and biophysical research communications* **171**, 162–168 (1990).
 114. Holwerda, S. & de Laat, W. Chromatin loops, gene positioning, and gene expression. *Front. Genet.* **3**, 217 (2012).
 115. Georgopoulos, K., van den Elsen, P., Bier, E., Maxam, A. & Terhorst, C. A T cell-specific enhancer is located in a DNase I-hypersensitive area at the 3' end of the CD3-delta gene. *The EMBO Journal* **7**, 2401–2407 (1988).
 116. Kingsley, P. D. *et al.* Ontogeny of erythroid gene expression. *Blood* **121**, e5–e13 (2013).
 117. Flygare, J., Rayon Estrada, V., Shin, C., Gupta, S. & Lodish, H. F. HIF1 α synergizes with glucocorticoids to promote BFU-E progenitor self-renewal. *Blood* **117**, 3435–3444 (2011).
 118. Socolovsky, M. *et al.* Ineffective erythropoiesis in Stat5a(-/-)5b(-/-) mice due to decreased survival of early erythroblasts. *Blood* **98**, 3261–3273 (2001).
 119. Creighton, M. P. & Cheng, A. W. Histone H3K27ac separates active from poised enhancers and predicts developmental state. in (2010). doi:10.1073/pnas.1016071107/-/DCSupplemental/st01.xlsx
 120. Kaikkonen, M. U. *et al.* Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. *Molecular Cell*

- 51, 310–325 (2013).
121. Prescott, S. L. *et al.* Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell* 1–17 (2015). doi:10.1016/j.cell.2015.08.036
122. Hughes, J. R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Publishing Group* **46**, 205–212 (2014).
123. Zeuner, A. *et al.* The Notch2-Jagged1 interaction mediates stem cell factor signaling in erythropoiesis. *Cell Death Differ.* **18**, 371–380 (2011).
124. Silva, M. *et al.* Erythropoietin can promote erythroid progenitor survival by repressing apoptosis through Bcl-X(L) and Bcl-2. *Blood* **88**, 1576–1582 (1996).
125. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research* **42**, D142–D147 (2013).
126. Sebastian, A. & Contreras-Moreira, B. footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics* **30**, 258–265 (2014).
127. Stormo, G. D., SCHNEIDER, T. D., GOLD, L. & EHRENFEUCHT, A. Use of the Perceptron Algorithm to Distinguish Translational Initiation Sites in Escherichia-Coli. *Nucleic Acids Research* **10**, 2997–3011 (1982).
128. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* **18**, 6097–6100 (1990).
129. Lehmann, O. J., Sowden, J. C., Carlsson, P., Jordan, T. & Bhattacharya, S. S. Fox's in development and disease. *Trends in Genetics* **19**, 339–344 (2003).
130. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).
131. Georges, A. B., Benayoun, B. A., Caburet, S. & Veitia, R. A. Generic binding sites, generic DNA-binding domains: where does specific promoter recognition come from? *The FASEB Journal* **24**, 346–356 (2010).
132. Nakagawa, S. & Gisselbrecht, S. S. DNA-binding specificity changes in the evolution of forkhead transcription factors. in (2013). doi:10.1073/pnas.1310430110/-/DCSupplemental
133. Marinkovic, D. *et al.* Foxo3 is required for the regulation of oxidative stress in erythropoiesis. *J. Clin. Invest.* **117**, 2133–2144 (2007).
134. Isern, J. *et al.* Single-lineage transcriptome analysis reveals key regulatory pathways in primitive erythroid progenitors in the mouse embryo. *Blood* **117**, 4924–4934 (2011).
135. Kusy, S. *et al.* NKX3.1 is a direct TAL1 target gene that mediates proliferation of TAL1-expressing human T cell acute lymphoblastic leukemia. *J. Exp. Med.* **207**, 2141–2156 (2010).
136. Siddharthan, R. Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix. *PLoS ONE* **5**, e9722 (2010).
137. Mathelier, A. & Wasserman, W. W. The Next Generation of Transcription Factor Binding Site Prediction. *PLoS Comput. Biol.* **9**, e1003214 (2013).
138. Szigeti, A. *et al.* Induction of necrotic cell death and mitochondrial permeabilization by heme binding protein 2/SOUL. *FEBS Letters* **580**, 6447–6454 (2006).

139. Pang, C. J. *et al.* Kruppel-Like Factor 1 (KLF1), KLF2, and Myc Control a Regulatory Network Essential for Embryonic Erythropoiesis. *Molecular and Cellular Biology* **32**, 2628–2644 (2012).
140. Zhou, D., Liu, K., Sun, C.-W., Pawlik, K. M. & Townes, T. M. KLF1 regulates BCL11A expression and γ - to β -globin gene switching. *Nature Publishing Group* **42**, 754–756 (2010).
141. Kadri, Z. *et al.* Direct Binding of pRb/E2F-2 to GATA-1 Regulates Maturation and Terminal Cell Division during Erythropoiesis. *PLoS Biol* **7**, e1000123–15 (2009).
142. Delgado, M. D., Hyland, K. M., Lerga, A. & Richard, C. Max and inhibitory c-Myc mutants induce erythroid differentiation and resistance to apoptosis in human myeloid leukemia cells. *Oncogene* (1997).
143. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246–1250 (2009).
144. Zhang, K. *et al.* Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Methods* **6**, 613–618 (2009).
145. ZHANG, Q. Y., Rombel, I., REDDY, G. N., GANG, J. B. & Shen, C. Functional Roles of in-Vivo Footprinted Dna Motifs Within an Alpha-Globin Enhancer - Erythroid Lineage and Developmental Stage Specificities. *J. Biol. Chem.* **270**, 8501–8505 (1995).
146. De Gobbi, M. *et al.* Tissue-specific histone modification and transcription factor binding in α globin gene expression. *Blood* **110**, 4503–4510 (2007).
147. Lower, K. M. *et al.* Analysis of Sequence Variation Underlying Tissue-specific Transcription Factor Binding and Gene Expression. *Human Mutation* **34**, 1140–1148 (2013).
148. Harteveld, C. L., Muglia, M., Passarino, G., Kielman, M. F. & Bernini, L. F. Genetic polymorphism of the major regulatory element HS-40 upstream of the human α -globin gene cluster. *Br. J. Haematol.* **119**, 848–854 (2002).
149. De Gobbi, M. *et al.* A Regulatory SNP Causes a Human Genetic Disease by Creating a New Transcriptional Promoter. *Science* **312**, 1215–1217 (2006).
150. Zhang, X., Cowper-Salari, R., Bailey, S. D., Moore, J. H. & Lupien, M. Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. *Genome Research* **22**, 1437–1446 (2012).
151. Bauer, D. E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253–257 (2013).
152. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* **373**, 895–907 (2015).
153. Higgs, D. R., Vernimmen, D. & Wood, B. Long-range regulation of alpha-globin gene expression. *Adv. Genet.* **61**, 143–173 (2008).
154. Lahlil, R., Lécuyer, E., Herblot, S. & Hoang, T. SCL assembles a multifactorial complex that determines glycophorin A expression. *Molecular and Cellular Biology* **24**, 1439–1452 (2004).
155. Vyas, P. *et al.* Different sequence requirements for expression in erythroid and megakaryocytic cells within a regulatory element upstream of the GATA-1 gene. *Development* **126**, 2799–2811 (1999).
156. Suck, D. DNA recognition by DNase I. *Journal of Molecular Recognition* **7**, 65–70 (1994).
157. Reznikoff, W. S. Tn5 as a model for understanding DNA transposition.

- Molecular Microbiology* **47**, 1199–1206 (2003).
158. Ong, C.-T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nature Publishing Group* **15**, 234–246 (2014).
 159. Ason, B. & Reznikoff, W. S. DNA Sequence Bias During Tn5 Transposition. *J. Mol. Biol.* **335**, 1213–1225 (2004).
 160. Lazarovici, A., Zhou, T. & Shafer, A. Probing DNA shape and methylation state on a genomic scale with DNase I. in (2013). doi:10.1073/pnas.1216822110/-/DCSupplemental
 161. Noordermeer, D. & de Laat, W. Joining the loops: β -Globin gene regulation. *IUBMB Life* **60**, 824–833 (2008).
 162. Xu, J. *et al.* Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. *Genes Dev.* **23**, 2824–2838 (2009).
 163. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nature Publishing Group* **16**, 144–154 (2015).
 164. Pop, R., Shearstone, J. R., Shen, Q., Liu, Y. & Hallstrom, K. A key commitment step in erythropoiesis is synchronized with the cell cycle clock through mutual inhibition between PU. 1 and S-phase progression. *PLoS Biol* (2010). doi:10.1371/journal.pbio.1000484.g007
 165. Cirillo, L. A. *et al.* Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *The EMBO Journal* **17**, 244–254 (1998).
 166. Guertin, M. J. & Lis, J. T. Mechanisms by which transcription factors gain access to target sequence elements in chromatin. *Current Opinion in Genetics & Development* **23**, 116–123 (2013).
 167. Lupien, M. *et al.* FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription. *Cell* **132**, 958–970 (2008).
 168. Magnani, L., Eeckhoutte, J. & Lupien, M. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends in Genetics* **27**, 465–474 (2011).
 169. Hashimoto, T. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology* **32**, 171–178 (2014).
 170. Wu, W. *et al.* Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Research* **21**, 1659–1671 (2011).
 171. Rohs, R. *et al.* Origins of Specificity in Protein-DNA Recognition. *Annu. Rev. Biochem.* **79**, 233–269 (2010).
 172. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* **39**, 381–399 (2014).
 173. Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**, 730–732 (2007).
 174. Weirauch, M. T. & Hughes, T. R. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends in Genetics* **26**, 66–74 (2010).
 175. Jackson, P. D. & Felsenfeld, G. A method for mapping intranuclear protein-DNA interactions and its application to a nuclease hypersensitive site. *PNAS* **82**, 2296–2300 (1985).
 176. McNally, J. G., Müller, W. G., Walker, D., Wolford, R. & Hager, G. L. The

- glucocorticoid receptor: rapid exchange with regulatory sites in living cells. *Science* **287**, 1262–1265 (2000).
177. Voss, T. C. & Hager, G. L. Dynamic regulation of transcriptional states by chromatin and transcription factors. 1–13 (2013). doi:10.1038/nrg3623
 178. Sung, M.-H., Guertin, M. J., Baek, S. & Hager, G. L. DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence. *Molecular Cell* **56**, 275–285 (2014).
 179. Lickwar, C. R., Mueller, F., Hanlon, S. E., McNally, J. G. & Lieb, J. D. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* **484**, 251–255 (2012).
 180. Sharp, Z. D. *et al.* Estrogen-receptor- α exchange and chromatin dynamics are ligand- and domain-dependent. *J Cell Sci* **119**, 4101–4116 (2006).
 181. Bosisio, D. *et al.* A hyper-dynamic equilibrium between promoter-bound and nucleoplasmic dimers controls NF- κ B-dependent gene activity. *The EMBO Journal* **25**, 798–810 (2006).
 182. Mazza, D., Abernathy, A., Golob, N., Morisaki, T. & McNally, J. G. A benchmark for chromatin binding measurements in live cells. *Nucleic Acids Research* **40**, gks701–e119 (2012).
 183. Malnou, C. E. *et al.* Heterodimerization with different Jun proteins controls c-Fos intranuclear dynamics and distribution. *J. Biol. Chem.* **285**, 6552–6562 (2010).
 184. Takahashi, M. & Saitou, N. Identification and Characterization of Lineage-Specific Highly Conserved Noncoding Sequences in Mammalian Genomes. *Genome Biology and Evolution* **4**, 641–657 (2012).
 185. Hammar, P. *et al.* The lac Repressor Displays Facilitated Diffusion in Living Cells. *Science* **336**, 1595–1598 (2012).
 186. Chen, C.-Y., Chang, I.-S., Hsiung, C. A. & Wasserman, W. W. On the identification of potential regulatory variants within genome wide association candidate SNP sets. *BMC Med Genomics* **7**, 34 (2014).
 187. Teytelman, L., Thurtle, D. M. & Rine, J. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. in (2013). doi:10.1073/pnas.1316064110/-DCSupplemental
 188. Erceg, J. *et al.* Subtle Changes in Motif Positioning Cause Tissue-Specific Effects on Robustness of an Enhancer's Activity. *PLoS Genet* **10**, e1004060–16 (2014).
 189. Swanson, C. I., Evans, N. C. & Barolo, S. Structural Rules and Complex Regulatory Circuitry Constrain Expression of a Notch- and EGFR-Regulated Eye Enhancer. *Developmental Cell* **18**, 359–370 (2010).
 190. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
 191. Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007–1012 (2014).
 192. Mendenhall, E. M. *et al.* Locus-specific editing of histone modifications at endogenous enhancers. *Nature Biotechnology* 1–5 (2013). doi:10.1038/nbt.2701
 193. Yáñez-Cuna, J. O., Kvon, E. Z. & Stark, A. Deciphering the transcriptional cis-regulatory code. *Trends in Genetics* **29**, 11–22 (2013).
 194. McCarthy, M. I. GENOMIC MEDICINE Genomics, Type 2 Diabetes, and

- Obesity. *N Engl J Med* **363**, 2339–2350 (2010).
195. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology* **32**, 347–355 (2014).
 196. Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nature Publishing Group* **46**, 61–64 (2013).
 197. Andrews, N. C., Erdjument-Bromage, H. & Davidson, M. B. Erythroid transcription factor NF-E2 is a haematopoietic-specific basic-leucine zipper protein. *Nature* **362**, 722–728 (1993).
 198. Ney, P. A. *et al.* Purification of the human NF-E2 complex: cDNA cloning of the hematopoietic cell-specific subunit and evidence for an associated partner. *Molecular and Cellular Biology* **13**, 5604–5612 (1993).
 199. Loyd, M. R., Okamoto, Y., Randall, M. S. & Ney, P. A. Role of AP1/NFE2 binding sites in endogenous α -globin gene transcription. *Blood* **102**, 4223–4228 (2003).
 200. Kurita, R. *et al.* Establishment of Immortalized Human Erythroid Progenitor Cell Lines Able to Produce Enucleated Red Blood Cells. *PLoS ONE* **8**, e59890–15 (2013).