Elucidating the *cis*-regulatory logic of *Runx1* during developmental haematopoiesis

Dominic David Gregory Owens

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019

In loving memory of Dr. S. M. Owens

Scientist, Teacher, and Mother.

Contents

List of Abbreviations

3C Chromosome conformation capture.

AGM Aorta-gonads-mesonephros.

bp base pair.

Cas9 Clustered regularly interspaced short palindromic repeat associated protein 9.

CGI CpG island.

cHE Competent haemogenic endothelium.

ChIP Chromatin immunoprecipitation.

ChIP-seq Chromatin immunoprecipitation with high-throughput sequencing.

CRISPR Clustered regularly interspaced short palindromic repeat.

CTCF CCCTC-binding factor.

DSB DNA double strand break.

E Embryonic day.

EB Embryoid body.

EHT Endothelial-to-haematopoietic transition.

EMP Erythro-myeloid progenitor.

EMSA Electrophoretic mobility shift assay.

eRNA Enhancer RNA.

HE Haemogenic endothelium.

hiPSC Locus control region.

hiPSC Human induced pluripotent stem cells.

HP Haematopoietic progenitor cells.

HR Homologous recombination.

HSC Haematopoietic stem cell.

HSPC Haematopoietic stem and progenitor cells.

Inr Initiator.

kb kilobase pair.

LD Larger deletion.

LEF Loop-extruding factor.

LR Long-range.

Mb megabase pair.

mESC Mouse embryonic stem cells.

MMEJ Microhomology-mediated end joining.

MR Medium-range.

NG Capture-C Next Generation Capture-C.

NHEJ Non-homologous end joining.

P1 *Runx1* distal promoter.

P2 *Runx1* proximal promoter.

PAS Para-aortic-splanchnopleura.

PCR Polymerase chain reaction.

Pol II RNA Polymerase II.

qPCR Quantitative polymerase chain reaction.

qRT-PCR Quantitative reverse transcription polymerase chain reaction.

SE Super enhancer.

sHE Specifying haemogenic endothelium.

siRNA Short interfering RNA.

SMC structural maintenance of chromosome.

SR Short-range.

SSA Single strand annealing.

TCT polypyrimidine initiator.

TF Transcription factor.

TFBS Transcription factor binding site.

TSS Transcription start site.

UCSC University of California, Santa Cruz.

List of Figures

List of Tables

Acknowledgements

This work was only possible thanks to all the people who shared their time, help, patience, and advice with me.

First, I would like to thank my supervisor and mentor Prof. Marella de Bruijn. Despite her, at times, full-on schedule, she always found time to discuss exciting (and puzzling) results with me, and to read my work. Much of the improvements in my scientific thinking and writing I owe to her.

I also owe so much to my lab mates. Dr. Vincent Frontera, to whom I am eternally grateful for his early morning optimistic experimental plans, last minute FACS disaster rescues, late night sorts, and emergency coffee breaks. Joe Harman, for convincing me that computers will take over the world and for always being around when R got the better of me. Dr. Christina Rode, for showing me that science and art are not always two separate things. Dr. Emanuele Azzoni, whose curiosity, optimism, and scientific rigour inspired me on many occasions. Dr. Lucas Greder, for teaching me so many laboratory skills, even when he was too busy. I am especially grateful to Akin Bucakci, my first student in the lab. He is a powerhouse at the bench and remains a great friend.

I am incredibly grateful to my supervisor Prof. Jim Hughes, and his lab. He inspired me to care about the structure of the genome and gave me great ideas for experiments. The first member of the Hughes lab to mention is Dr. Damien Downes. He is a great scientist who taught me Capture-C, and how to enjoy doing it. Dr. Marieke Oudelaar provided so much useful advice and always gave great feedback on my written work. Jelena Telenius is a formidable bioinformatician who wrote several data analysis pipelines that were incredibly useful for this thesis. Ron Schwessinger is another inspirational bioinformatician, who gave up a lot of time to help me with DNaseI footprinting.

Outside the de Bruijn and Hughes labs, several members of the MHU and WIMM have been incredibly influential on me. Danuta Jeziorska was always enthusiastic about my work and generous with her time. Yavor Bozhilov gave me expert advice regarding genome editing. Philip Hublitz also provided expert genome editing advice, and someone to talk to at the nanodrop.

My adoptive lab was that of Prof. Tom Milne. Tom was great to talk to in the corridor and is an inspirational scientist and group leader. I thoroughly enjoyed going to the Milne lab meetings for free coffee and interesting discussions, despite the early starts. Nick Crump and Laura Godfrey especially provided frequent advice and reagents which were gratefully received.

Lydia Teboul was a fantastic collaborator on our CRISPR project. She is a great thinker and very knowledgeable. Lydia, and members of her lab (Adam Caulder and Alasdair Allan) were all essential to the eventual publication of part of this thesis (sections 4.2.6 to 4.2.9). When we saw unexpected results in our CRISPR experiments

it felt like a disaster, but thankfully, with the help of Lydia and her group it formed the basis of my first publication (Owens et al., 2019).

I could never thank my family enough. My mother, Dr. Suzie Owens, and father, Prof. Nick Owens, were the first scientists I ever knew. They always supported me in my education and instilled a wonder in nature that I carry with me to this day. Dr. Benjamin Owens has offered so much help and advice, and always given me someone to look up to over the years. He is the reason I came to Oxford in the first place. Matthew Owens inspired me to keep physically active and I have great memories of his regular visits to Oxford. Dr. Timothy Owens helped keep me sane by always welcoming me to his home in St. Andrews.

Finally, I will always be indebted to Sofia Deleniv; my best friend, colleague, and partner. I could not have done this without her.

List of publications

Dominic D G Owens, Adam Caulder, Vincent Frontera, Joe R Harman, Alasdair J Allan, Akin Bucakci, Lucas Greder, Gemma F Codner, Philip Hublitz, Peter J McHugh, Lydia Teboul, Marella F T R de Bruijn. **Microhomologies are prevalent at Cas9-induced larger deletions**, *Nucleic Acids Research*, Volume 47, Issue 14, 22 August 2019, Pages 7402–7417, https://doi.org/10.1093/nar/gkz459

Abstract

The study of dynamic gene expression during development utilises model differentiation systems. The haematopoietic system is one of the archetypal differentiation systems and many model gene loci used to shed light on basic mechanisms of transcriptional regulation are important haematopoietic genes. In this project I set out to uncover transcriptional regulatory mechanisms of the gene *Runx1*. *Runx1* is a transcription factor important for the development and function of the haematopoietic system. In particular, *Runx1* is required for haematopoietic stem cell (HSC) generation during embryonic development. HSCs are born during an endothelial-to-haematopoietic transition (EHT). *Runx1* is expressed during, and is absolutely required for, EHT and the birth of HSCs. Previous work in the lab established that *Runx1* exhibits complex transcriptional regulation during EHT; *Runx1* has two alternative promoters and several haematopoietic enhancers have been identified.

Using cutting-edge chromatin assays I interrogated *cis*-interactions within the transcriptional regulatory domain of *Runx1* when it was expressed at low and high levels. Capture-C showed that interactions between *Runx1* promoters and the haematopoietic enhancers previously identified were more frequent in haematopoietic cells compared to undifferentiated mouse embryonic stem cells (mESCs). Increased *Runx1* transcription was associated with an apparent increase in the hallmarks of loop extrusion, suggesting that the two might be causally related.

Next, I set out to investigate whether functional redundancy exists between the haematopoietic *Runx1* enhancers. Distinct upstream regulators of the enhancers that were identified suggested that they might not have entirely overlapping functions. While using CRISPR/Cas9 in mESCs to delete endogenous *Runx1* enhancers and test their functional requirements, I uncovered interesting caveats about common genome engineering practices. Functional characterisation of enhancer-deleted mESCs differentiated *in vitro* to haematopoietic cells indicated that partial enhancer redundancy may exist.

Finally, I investigated a potential mechanism of *Runx1* alternative promoter choice involving CTCF/cohesin. I identified tissue-specific and ubiquitous CTCF binding sites in the *Runx1* domain that were capable of acting as insulators *in vitro*. Cell type-specific binding of CTCF upstream of each of the *Runx1* promoters was correlated to DNA demethylation and promoter activity. CTCF binding was seen upstream of a significant proportion of active promoters genome-wide, suggesting that modulating CTCF binding close to promoters, possibly via DNA methylation, may be a widespread transcriptional regulatory mechanism.

1. Introduction

All cells in the human body share the same genetic material. Gene expression is the tightly regulated process by which cells interpret this genetic code. Differential gene expression facilitates a dizzying array of cellular phenotypes and gene regulation is the process that defines the set of expressed genes in a given cell at a given time. Gene expression is dynamic in space and time during embryogenesis—impacting the ultimate form and function of the organism. Moreover, when gene regulation goes awry, complex disease results (Lee and Young, 2013). Since the completion of the Human Genome Project almost two decades ago (Venter et al., 2001), the cost of sequencing a genome has gone from millions to hundreds of dollars, and the time it takes, from years to mere hours. It has become increasingly clear that the regulatory information that facilitates regulated gene expression is located in non-coding DNA. In this postgenomic era, a major challenge remains—how do we interpret this code? At least in principle, this must be possible; the biological machinery inside living cells achieves it correctly, almost without fail.

1.1 Model systems used to study gene regulation

One of the archetypal systems for studying gene regulation is the haematopoietic system. Haematopoietic stem cells (HSCs) reside at the top of a cellular hierarchy ultimately generating all the functionally diverse blood cell types including red blood cells (erythrocytes), platelets, and white blood cells (Rieger and Schroeder, 2012) (Figure 1.1). Cellular differentiation to these different lineages requires tight regulation of gene expression. Haematopoietic cells are abundant, readily accessible, and can often be differentiated or activated *in vitro* accompanied by dynamic gene expression changes. Together this makes them an ideal system for studying mechanisms of gene regulation.

A particularly heavily studied haematopoietic cellular differentiation model system is erythropoiesis (Hattangadi et al., 2011). Erythropoiesis begins with HSCs and ultimately generates millions of red blood cells per second in adult humans. These red cells transport oxygen, bound by haemoglobin. The high levels of haemoglobin generated in these cells, along with the relative ease of isolating them have made the study of erythroid-specific α- and β-globin transcription crucial to establishing our current view of the dynamics of tissue-specific gene regulation (Noordermeer and de Laat, 2008; Higgs et al., 2008). Other haematopoietic cellular differentiation systems have also been used to study dynamic gene regulation, including the differentiation of naïve T cells to T helper cells (Zhu and Paul, 2008) and the maturation of B lineage cells (Ebert and Busslinger, 2015). Haematopoiesis is probably the best studied cellular differentiation system, but other model systems such as the developing limb bud (Zuniga and Zeller, 2014) have also yielded key insights. Each developmental model system has its own pros and cons, and they all provide a window onto dynamic gene expression changes during cellular differentiation.

1.1.1 Developmental origins of the adult haematopoietic system

The development of the haematopoietic system has been extensively studied in mice. Blood cell development occurs in three waves during mouse embryogenesis (Figure 1.2 a). During the first wave primitive erythrocytes are generated in the blood islands of the murine yolk sac between E7.0 and E7.5, along with primitive macrophages, and megakaryocytes (Moore and Metcalf, 1970; Palis et al., 1999). The second wave also takes place in the yolk sac, and from E8.25 generates erythro-myeloid progenitors (EMPs) with 'definitive' erythroid, myeloid, megakaryocyte potential, and limited lymphoid potential (McGrath and Palis, 2015; Palis et al., 1999). Definitive HSCs are born during the third wave of haematopoiesis in the aorta-gonad-mesonephros region (AGM) of the embryo proper at E11.5 (Müller et al., 1994; Medvinsky and Dzierzak, 1996; de Bruijn et al., 2000).

1.1.2 *In vitro* models of embryonic haematopoiesis

Efforts have also gone into developing *in vitro* models of embryonic haematopoesis. *In vitro* models traditionally relied on mouse embryonic stem cells (mESCs) to study the stages of differentiation from pluripotent cells to haematopoietic cells (Lacaud et al., 2004). More recently, the differentiation of human induced pluripotent stem cells (hiPSCs, Takahashi et al. 2007; Yu et al. 2007) has provided insights into human developmental haematopoiesis which has traditionally been recalcitrant to study *in vivo* (Ditadi et al., 2017). Various differentiation strategies exist—some utilising serum or stromal cell layers to provide the signals necessary to direct differentiation (Ledran et al., 2008), and others using chemically defined medium (Uenishi et al., 2014). A greater degree of control can be achieved using chemically defined medium with the addition of specific cytokines at specified developmental timepoints to direct differentiation to desired lineages (Irion et al., 2008; Ditadi et al., 2017). The generation of HSCs from hiPSCs has great therapeutic potential through offering cell transplantation therapies for various diseases including auto-immune diseases and leukaemia. Despite great advances in the generation of HSCs *in vitro* (Lis et al., 2017; Sugimura et al., 2017), this is currently not feasible without using exogenous transgenes. Increasing our knowledge of the development of HSCs *in vivo* using model systems like mouse developmental haematopoiesis may offer new insights into the *de novo* generation of HSCs *in vitro*.

1.1.3 From endothelium to haematopoietic stem cells

In the AGM, HSCs are generated by a process of endothelial-to-haematopoietic transition (EHT). *In vivo*, detailed anatomical observations suggested that haemogenic endothelial (HE) cells appeared to be undergoing EHT and budding off into the aorta (Jaffredo et al., 1998). The process of EHT was observed directly *in vitro* through live imaging of mESC cultures (Eilken et al., 2009; Lancrin et al., 2009). Elegant live imaging studies in mouse AGM explants and zebrafish embryos showed HE cells budding from the wall of the dorsal aorta (Bertrand et al., 2010; Kissa and Herbomel, 2010; Boisset et al., 2010). In the mouse these budding cells form HSC-containing intra-aortic haematopoietic clusters that co-express endothelial (e.g. CD31, VE-cadherin) and haematopoietic (e.g. CD41, CD45) markers, in line with an endothelial origin

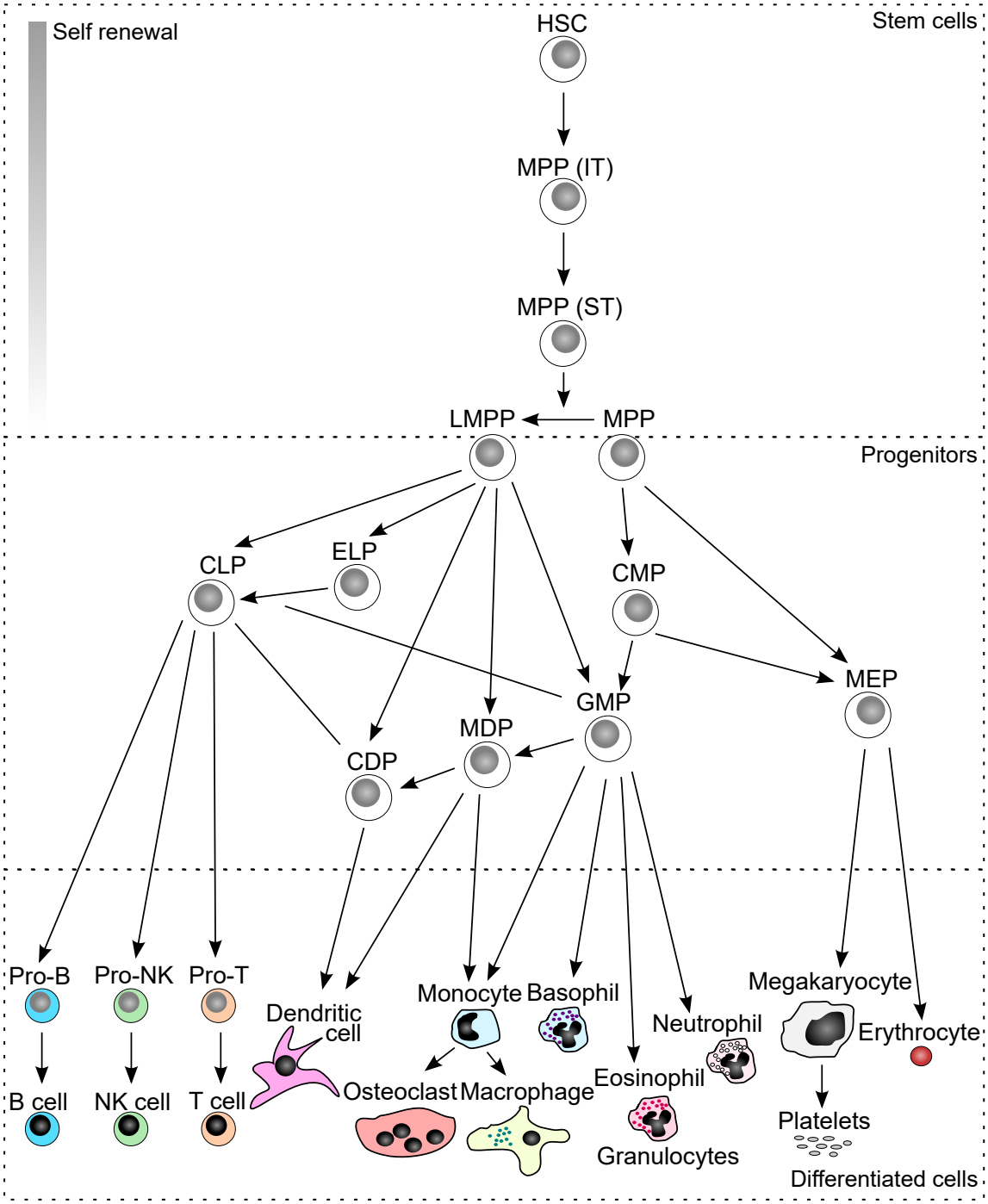


Figure 1.1 – Adult haematopoietic stem cell differentiation hierarchy. Long-term self renewing haematopoietic stem cells (LT-HSCs) reside at the top of a cellular hierachy ultimately leading to the generation of all mature blood cell lineages. LT-HSCs first differentiate into multipotent progenitors (MPP) with intermediate-term (IT) or short-term (ST) self-renewal potential. Differentiation then proceeds down various different progenitor cell types including lymphoid-primed MPP (LMPP), early lymphoid progenitor (ELP), common lymphoid progenitor (CLP), common myeloid progenitor (CMP), granulocyte-macrophage progenitor (GMP), megakaryocyte-erythroid progenitor (MEP), common dendritic progenitor (CDP), monocyte-dendritic progenitor (MDP). Well-described routes of differentiation are shown as thick arrows and more recently described differentiation pathways are shown as thin arrows. This model is not absolute and is frequently being updated based on ongoing research in the field. Adapted from Rieger and Schroeder 2012.

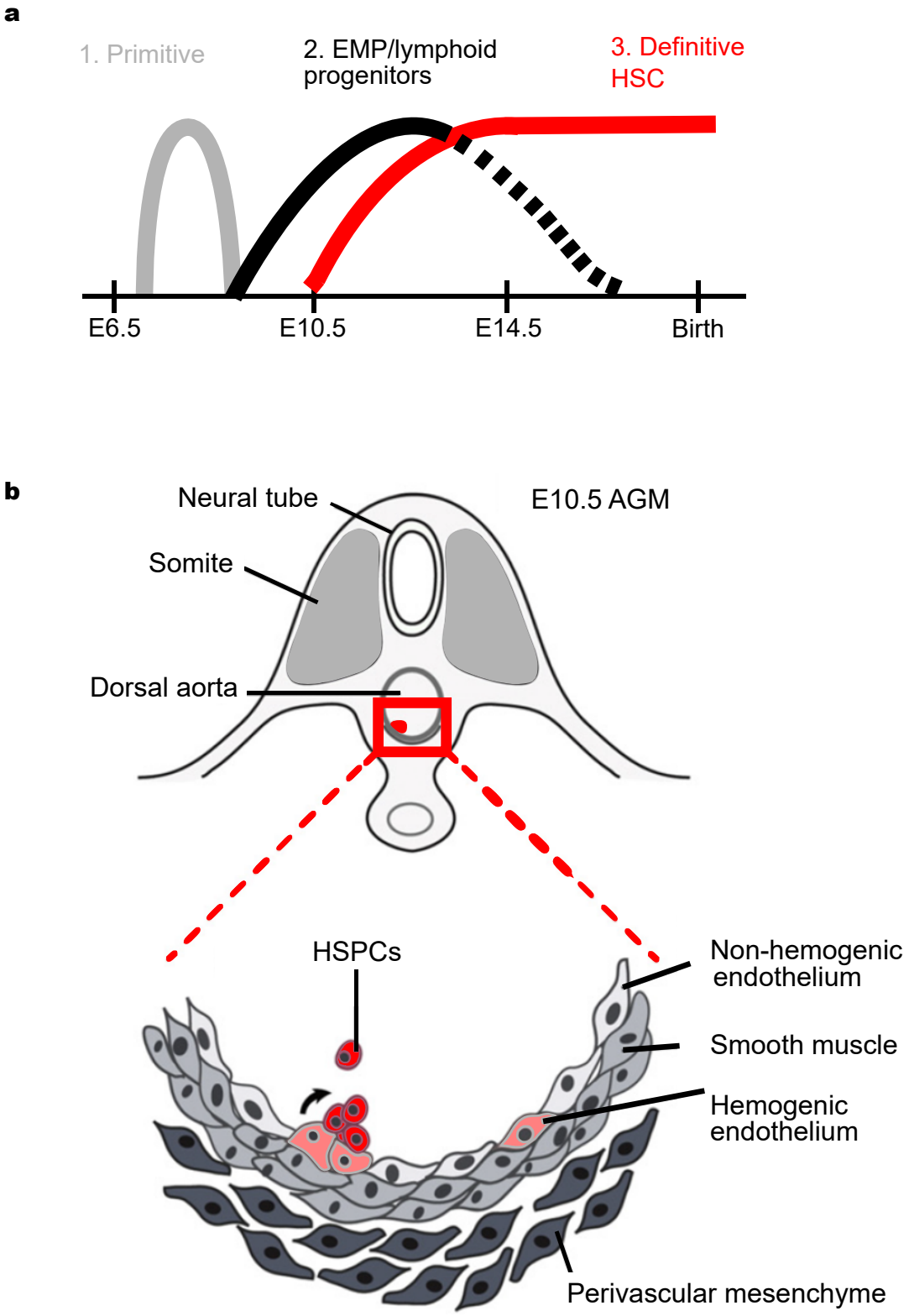


Figure 1.2 – The origin of definitive haematopoietic stem cells. a) The three partially overlapping but distinct waves of haematopoietic cell development during ontogeny. Definitive haematopoietic stem cells (HSCs) are generated in the dorsal aorta from haemogenic endothelium (HE). b) Schematic showing a cross section through the aorta-gonads-mesonephros (AGM) region of a E10.5 mouse embryo. Haematopoietic stem and progenitor cells (HSPCs, shown in red) emerge from HE cells (pink) by a process of endothelial-to-haematopoietic transition (EHT), forming clusters of cells that ultimately detach and enter circulation. Both as depicted by Swiers et al. 2013b.

of HSCs (Swiers et al., 2013b). Work in our laboratory showed that HE cells gain haematopoietic potential while still embedded in the endothelial wall, indicative of early molecular changes underlying the differentiation from endothelium to blood (Swiers et al., 2013a).

Several factors have been implicated in the generation of haematopoietic stem and progenitor cells (HSPCs) including cell surface signalling receptors such as Flk1 (Shalaby et al., 1995), G-protein coupled receptors (Kartalaei et al., 2015; Zhang et al., 2015; Gao et al., 2013) and inflammatory cytokine receptors (Sawamiphak et al., 2014; Li et al., 2014a; Espín-Palazón et al., 2014). Notch1 signalling is also absolutely required for HSC specification (Kumano et al., 2003), but is downregulated during HSC maturation (Souilhol et al., 2016). Ultimately, signalling events at the cell surface induce genetic changes in the nucleus through influencing DNA binding transcription factors (TFs). Many TFs play crucial roles in the development of the haematopoietic system including Scl (Shivdasani et al., 1995), Gata2 (Tsai et al., 1994), Lmo2 (Warren et al., 1994), Meis1 (Azcoitia et al., 2005), and E-twenty-six (ETS) factors Fli1 (Abedin et al., 2014), Etv2 (Lee et al., 2008) and Etv6 (Wang et al., 1997). These TFs are important for specifying mesoderm, endothelium and/or the haematopoietic lineage. As such, many of them are also important for the blood formation waves preceding HSC emergence.

1.2 The origins of the Runx gene family

Runx1 is a member of the Runt-related (*RUNX*) family of TFs (comprising of *RUNX1*, *RUNX2* and *RUNX3* in mammals) that are involved in various processes in development and disease (Levanon and Groner, 2004; de Bruijn and Dzierzak, 2017; Vimalraj et al., 2015; Lotem et al., 2017; Deltcheva and Nimmo, 2017; Mevel et al., 2019). The Runx gene family has ancient origins. Unicellular protozoans lack detectable Runx genes, with basal animals including Cnidaria, sponges, and *Caenorhabditis elegans* containing only a single Runx gene (Sullivan et al., 2008). Independent duplications are thought to have occurred in the protostome and deuterostome lineages (Sullivan et al., 2008), resulting in *Drosophila melanogaster* (a protostome) containing four distinct Runx genes, and mice and humans (both deuterostomes) containing three (Levanon and Groner, 2004).

Orthologous and paralogous Runx genes share several similarities. At the protein level, all Runx genes exhibit substantial homology in their DNA-binding (RD) domain (Sullivan et al., 2008). As such, RUNX factors recognise the same canonical motif '`YGYGGT`' (where `Y` is a pyrimide base `C/T`) (Levanon et al., 2001). Alongside homology at the protein level, Runx genes also share similarities in their genomic structure and location. In human and mouse, each of the *RUNX* genes resides within a paralogous region containing a *CLIC* gene and a *RCAN* gene (green bars, Figure 1.3). A similar association between Runx and Clic genes is also observable in other species including *Gallus gallus*, *Xenopus laevis*, and *Danio rerio* (Figure 1.3), further emphasising the tendency for this paralogous region to be conserved through evolution. Each of the paralogous *RUNX* genes reside within a larger block of synteny conserved in human-mouse-chicken (blue bars, Figure 1.3) and human-mouse-frog (orange bars, Figure 1.3) containing multiple other genes (Ahituv et al. 2005, Fig-

ure 1.3). A neighbouring gene desert is another conserved feature of the Runx gene family. *RUNX1* and *RUNX3* both reside next to a large gene desert, while *RUNX2* shows the hallmarks of this gene desert but with the *SUPT3H* gene residing within it (black bars, Figure 1.3). Together, the shared paralogous genes, shared synteny, and shared gene desert all suggest evolutionary constraints exist that maintains the structure of Runx genes and their wider genomic loci. It is plausible to hypothesise that this constraint is due to the presence of sequences important for regulating *RUNX* gene expression within the paralogous genes and/or within the neighbouring gene desert.

1.3 Runx1 and its role in generating haematopoietic stem cells

Runx1 is of special interest to the study of HSPC development due to its critical role in EHT. Runx1 null mouse embryos die around midgestation with foetal liver anaemia and central nervous system haemorrhages; they lack all blood cells apart from primitive erythrocytes (Okuda et al., 1996; Wang et al., 1996; North et al., 1999; Cai et al., 2000; Yokomizo et al., 2008). Both *in vitro* in mESC-derived haematopoiesis and *in vivo* in the mouse embryo, Runx1 is required for cells to undergo EHT (Lancrin et al., 2009; Chen et al., 2009). It is central to a complex gene regulatory network orchestrating EHT (Swiers et al., 2010). TFs downstream of the gene include Gfi1b (Thambyrajah et al., 2016), Gfi1 (Wilson et al., 2010b), Pu.1 (Okada et al., 1998), and Myb (Okada et al., 1998). Upstream regulators of *Runx1* in developmental haematopoiesis include Notch signalling and Gata2 (Robert-Moreno et al., 2005; Burns et al., 2005; Nottingham et al., 2007), Wnt/β-catenin signaling (Medina et al., 2016), Scl (Nottingham et al., 2007; Schütte et al., 2016; Pimanda et al., 2007), Ets factors (Nottingham et al., 2007; Schütte et al., 2016), and Bmp4 signaling acting through Smad1 (Pimanda et al., 2007). Changes in *Runx1* dosage affects haematopoiesis in the embryo (Cai et al., 2000; Wang et al., 1996; Lie-A-Ling et al., 2018) and its transcription is regulated dynamically during development (Nottingham et al., 2007; Bee et al., 2009b; Yzaguirre et al., 2017). Recent attempts to generate HSCs *de novo in vitro* using transgenes required the expression of *Runx1/RUNX1* (Lis et al., 2017; Sugimura et al., 2017) further highlighting the importance of the gene in HSC specification. A better understanding of endogenous *Runx1* regulation could inform future protocols for the *de novo* generation of HSCs *in vitro*. Below I summarise general principles of gene regulation and what is known about the transcriptional regulation of *Runx1*.

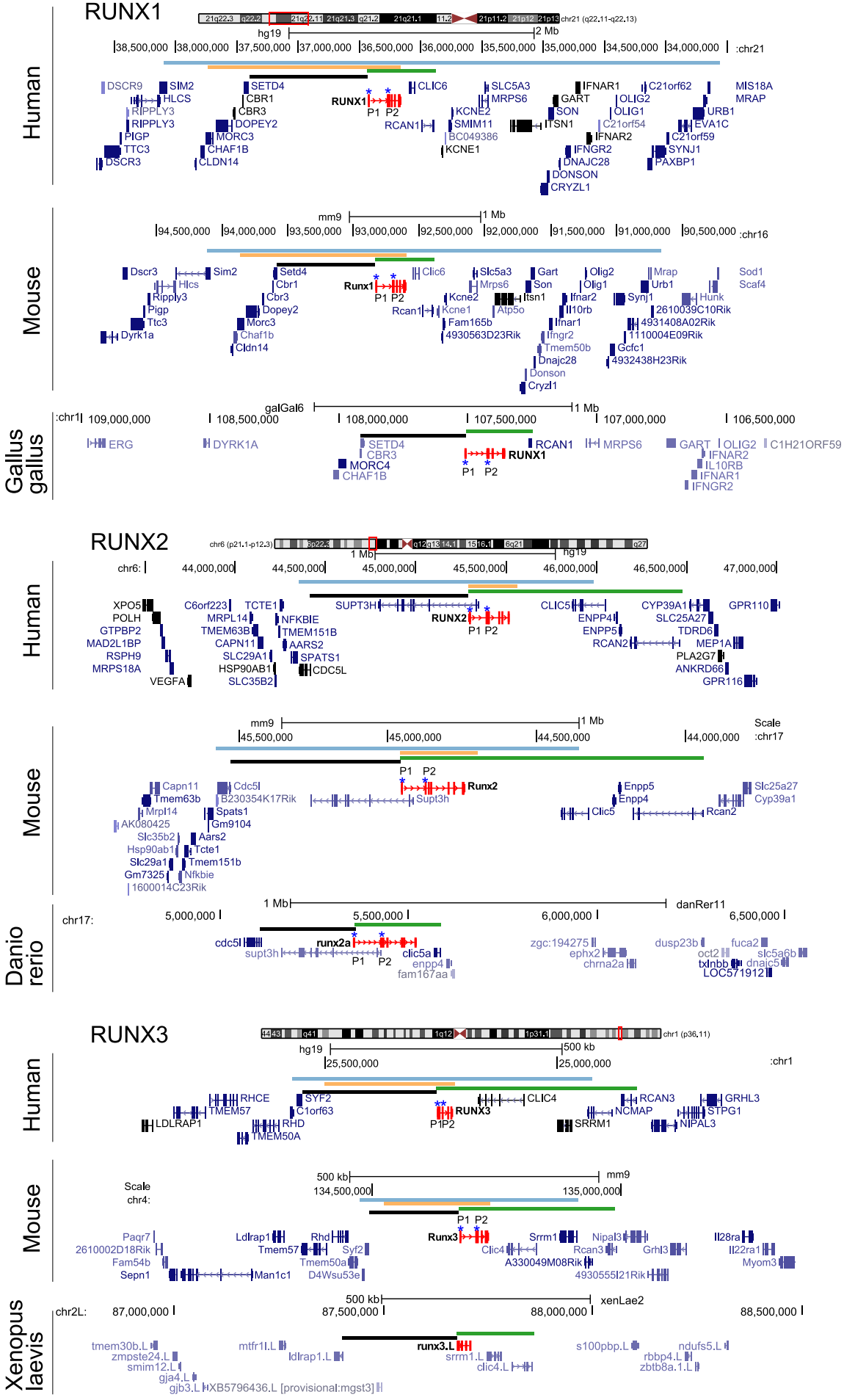


Figure 1.3 – The *RUNX* gene family in human and mouse. *Legend continued on next page.*

Figure 1.3 – *Legend continued from previous page.* Schematic depicting the genomic loci of *RUNX1*, *RUNX2*, and *RUNX3* in both human and in mouse, and Runx1 in chicken (*Gallus gallus*), runx2a in zebrafish (*Danio rerio*), and runx3 in frog (*Xenopus laevis*). The *RUNX* genes are highlighted in red. Note the two alternative promoters (labelled and highlighted with blue stars) that are a highly conserved feature of all *RUNX* genes. Previously identified conserved blocks of synteny in human-mouse-chicken (blue bars) and human-mouse-frog (orange bars) are shown (Ahituv et al., 2005). The paralogous region containing a Runx gene with a Clic and/or Rcan gene are indicated by green bars. Conserved gene deserts (or region reflective of one in the case of RUNX2) are indicated with a black bar. Gene depictions are from the UCSC browser with RefSeq gene annotations shown.

1.4 Gene regulation at the molecular level

Gene regulation is coordinated through the action of cell-type specific TFs that are activated by external or internal cues. These TFs cooperate with basal transcriptional machinery including RNA polymerase II (Pol II) and the pre-initiation complex to allow mRNA production (Lenhard et al., 2012). TFs elicit gene regulatory effects through binding at *cis*-regulatory elements of genes including promoters and enhancers.

1.4.1 Promoters

Core promoters are stretches of DNA up to a few hundred base pairs (bp) in length flanking the transcription start site (TSS) of a gene (Lenhard et al., 2012). Several classes of promoters exist, discriminated by their activities and nucleotide composition. In vertebrates, a common core promoter sequence motif is the TATA box with initiator (Inr) sequence (Lenhard et al., 2012; Danino et al., 2015). TATA box-containing promoters have a sharp and defined TSS and are associated with tissue-specific expression in adult tissues (Ponjavic et al., 2006; Carninci et al., 2006). Another subset of vertebrate promoters contain CpG islands (CGIs), that are ∼500-2000 bp regions with high GC content and are enriched for CpG dinucleotides (Deaton and Bird, 2011). CGI promoters have broader and less defined TSSs than TATA box promoters (Lenhard et al., 2012). However, the definition between TATA box and CGI promoters is not absolute. For example, polypyrimidine initiator (TCT motif) promoters contain a TATA box and overlap CGIs (Parry et al., 2010).

Developmentally regulated promoters are enriched for CGIs (Lenhard et al., 2012). They are frequently regulated by DNA methylation (Reik, 2007), and polycomb group proteins (Schuettengruber et al., 2007)—both important developmental regulators of transcription. Most active promoters are nucleosome depleted, marked by the histone modifications H3K4me3 and H3K27ac, and are DNaseI hypersensitive (Lenhard et al., 2012). Additionally, CGI promoters may be DNaseI hypersensitive and marked by both H3K4me3 and H3K27me3 when in a poised or bivalently marked state early during embryonic development (Lenhard et al., 2012).

1.4.1.1 *Runx1* promoters

The human *RUNX1* gene (also known as *AML1*) contains two alternative promoters (Miyoshi et al., 1995; Ghozi et al., 1996). This alternative-promoter structure is highly conserved between the orthologous genes in different species and within the orthologous *RUNX* genes (*RUNX1*, *RUNX2*, and *RUNX3*) in multiple species (Levanon and Groner, 2004) (Figure 1.3). The highly conserved alternative promoter structure suggests that it plays an important role in the regulation and/or function of *RUNX* genes.

Distinct functional roles for the *Runx1* P1 and P2 promoters have been described in developmental haematopoiesis (Sroczynska et al., 2009a; Bee et al., 2010). The P2 promoter lies within a large highly conserved CGI (Levanon and Groner, 2004) and is active first during development—at E7.5 in mouse primarily in primitive erythrocytes developing in the blood islands of the yolk sac (North et al., 1999; Telfer and Rothen-

berg, 2001; Lacaud et al., 2002; Bee et al., 2009b). In the para-aortic splanchnopeura (PAS), the precursor to the AGM, the P2 promoter becomes active at E8.5 (Bee et al., 2009b, 2010). The P1 promoter becomes active later than P2—not until after E9.5 in the PAS—and unlike P2 is CpG poor (Bee et al., 2009b, 2010). Both P1 and P2 promoters, however, contain non-canonical TATA box-like sequence motifs (Bee et al., 2009b, 2010). Both P1 and P2 activity is required for EHT to occur normally in the mouse embryo (Bee et al., 2010)—although P2 plays the more significant role (Sroczynska et al., 2009a; Bee et al., 2010). The *Runx1* P1 and P2 promoters are active to differing degrees in a variety of haematopoietic lineages (Telfer and Rothenberg, 2001). P1 is active in T-cells (Telfer and Rothenberg, 2001), and P1-derived Runx1c is required for normal megakaryopoiesis (Draper et al., 2017). P1-derived Runx1c has also been associated with the differentiation of HSPCs in hiPSC-derived *in vitro* differentiation cultures (Challen and Goodell, 2010; Navarro-Montero et al., 2017). Previous work established that the *Runx1* promoters alone do not confer haematopoietic specific gene expression (Ghozi et al., 1996; Bee et al., 2009b). In general, developmentally-regulated genes exhibit a higher level of transcriptional regulation by enhancers compared to other types of genes (Ernst et al., 2011). Therefore, *cis*-regulation by enhancers is likely to play a major role in the regulation of *Runx1* transcription.

1.4.2 Enhancers

Enhancers are key *cis*-regulatory elements that allow for complex tissue-specific gene expression. Together, the concerted action of sets of enhancers act to regulate the tissue-specific expression of important developmental genes, which in turn facilitate cellular differentiation and the specification of tissues (Buecker and Wysocka, 2012). The crucial role enhancers play is highlighted by the fact that the genomes of complex organisms like mammals contain vastly greater amounts of non-coding DNA (such as enhancers) than more simple organisms like *C. elegans* (Taft et al., 2007), while both sharing a similar number of protein coding genes.

Since their first identification decades ago, enhancers have been identified as important features of transcriptional regulatory mechanisms for a plethora of genes from flies to humans (Long et al., 2016). Enhancers are clusters of tissue-specific TF binding sequences that are able to integrate signals from multiple different upstream pathways (Buecker and Wysocka, 2012). Although the mechanisms of enhancer function at a distance are still being elucidated (Section 1.5), ultimately, enhancers bound by TFs activate nearby genes by facilitating Pol II-mediated gene transcription from neighbouring promoters (Ong and Corces, 2011). The first transcriptional enhancer identified was a stretch of SV40 viral DNA that was able to upregulate erythroid-specific β-globin gene transcription at a genomic distance and in an orientation-independent manner when transfected ectopically into cell lines (Banerji et al., 1981; Moreau et al., 1981). Alongside studying enhancers using artificial constructs, enhancers have also been extensively characterised at their endogenous loci. The first endogenous mammalian enhancer was identified in the immunoglobulin heavy chain locus that is expressed in a wide variety of immune cell types (Banerji and Schaffner, 1983; Gillies et al., 1983; Neuberger, 1983). Enhancers tested in episomal enhancer-reporter assays may exhibit different activities to enhancers in their endogenous loci as

chromatinisation of episomal plasmid constructs is unlikely to accurately recapitulate the endogenous locus (Arnold et al., 2013; Muerdter et al., 2018; Catarino and Stark, 2018). Enhancers can also drive the expression of reporter genes when integrated into the genome (Visel et al., 2008; Catarino and Stark, 2018). Random integration of enhancer-reporters may be favoured in open chromatin close to active enhancers or promoters which might over-estimate enhancer activities (Bushman, 2003; Myers et al., 2005; Akhtar et al., 2013; Catarino and Stark, 2018). These biases can be accounted for by examining multiple different transgenic embryos and by analysing multiple founder transgenic lines.

In recent years, enhancers have been identified genome-wide through high-throughput sequencing-based epigenetic assays (Maston et al., 2012). For example, active enhancers are considered to be enriched in histone modifications such H3K4me1 (Heintzman et al., 2007) and H3K27ac (Creyghton et al., 2010), DNaseI hypersensitivity (Crawford et al., 2006), p300 binding (Ghisletti et al., 2010) and tissue-specific TF binding (Maston et al., 2012). Enhancers can also be identified bioinformatically, by the presence of clusters of TF binding sites (TFBS) (Hallikas et al., 2006), and high levels of evolutionary conservation (Blanchette et al., 2006). Interestingly though, not all enhancers are evolutionarily conserved (Blow et al., 2010), such as the mouse specific α-globin enhancer, Rm (Hay et al., 2016).

1.4.2.1 *Runx1* haematopoietic enhancers

The first *Runx1* enhancer (denoted as +23, the distance in kb from the start codon of the first exon) was identified in our laboratory (Nottingham et al., 2007). It confers tissue-specific activity to a minimal promoter in a *Runx1* haematopoietic-specific pattern in mouse and zebrafish embryos *in vivo* (Figure 1.4 a, Nottingham et al. 2007; Ng et al. 2010). Since the identification of the +23 enhancer, several other haematopoietic *Runx1* enhancers have been identified, including -59, +3, +110, and +204 (Figure 1.4 a, Schütte et al. 2016). Together, this revealed that the activities of each of the individual *Runx1* enhancers recapitulates part of the overall *Runx1* expression pattern in the developing haematopoietic system. +23 enhancer-reporter expression is seen in all cells undergoing EHT and in all functional haematopoietic stem and progenitor cells (HSPCs) (Figure 1.4 b, Nottingham et al. 2007; Bee et al. 2010; Swiers et al. 2013a). The +110 enhancer drives expression later than +23—in a subset of HSPCs in intra-aortic cluster cells after undergoing EHT (Schütte et al. 2016 and de Bruijn lab unpublished results). The +204 enhancer drives expression in the HE cells prior and during EHT, and expression is reduced later in cluster cells (Schütte et al. 2016 and de Bruijn lab unpublished results). The -59 and +3 enhancers were also capable of driving reporter expression in haematopoietic tissues (Schütte et al., 2016), while the -328, -42/43 *cis*-elements did not drive reporter expression in haematopoietically active sites (Schütte et al., 2016).

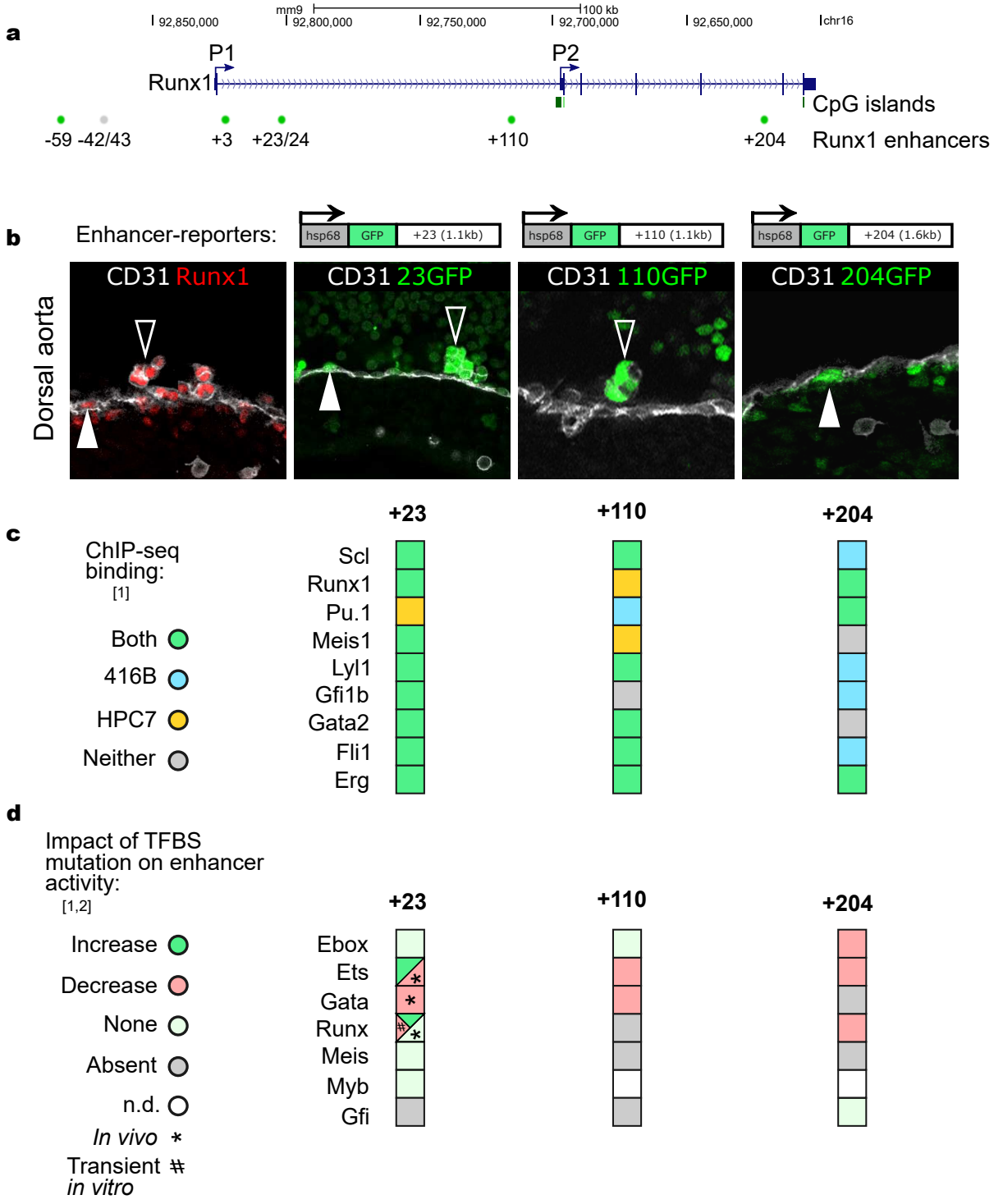


Figure 1.4 – *Runx1* enhancer activities and upstream transcription factor binding sites. a) Schematic of the *Runx1* gene spanning 224 kb in mouse on chromosome 16. Two alternative promoters (proximal P2 and distal P1) are indicated. Haematopoietic enhancers previously shown to drive reporter expression in haematopoietically active sites are indicated as green circles (named based on the distance in kb from the start codon of the first exon Nottingham et al. 2007; Bee et al. 2009b, 2010; Swiers et al. 2013a; Schütte et al. 2016). Putative enhancers that did not drive expression to haematopoietically active sites are shown as grey circles. CpG island annotations taken from the UCSC genome browser are shown. *Legend continued on next page.*

Figure 1.4 – *Legend continued from previous page.* b) Cross-sections through the mouse dorsal aorta in wild type and transgenic animals harbouring the *Runx1* enhancer-reporter transgenes indicated. Immunostaining of endothelial marker CD31 is shown in white, Runx1 is shown in red, and GFP is shown in green. Haemogenic endothelial (HE) cells expressing GFP under the control of the +23 and +204 enhancers are indicated by solid white arrow heads. Haematopoietic progenitor (HP) cells expressing GFP under the control of +23 and +110 enhancers are indicated by hollow white arrow heads. Imaging was performed by Christina Rode, Emanuele Azzoni, and Vincent Frontera. 23GFP line, Nottingham et al. 2007; Bee et al. 2010; 110GFP line (Andrew Jarrat and Vincent Frontera, unpublished data); 204GFP line (Christina Rode, unpublished data). c) Analysis of transcription factor binding to *Runx1* enhancers in haematopoietic cell lines ([1], Schütte et al. 2016). d) Upstream binding sites shown to regulate *Runx1* enhancers both *in vitro* ([1], Schütte et al. 2016) and *in vivo* ([2], Nottingham et al. 2007). # indicates where a discrepant result was shown in transient transgenic enhancer-reporter *in vitro* assays. * indicates where a result was shown *in vivo* in transient enhancer-reporter embryos.

1.4.3 Regulation of enhancer activity by *trans*-acting factors

A hallmark of active enhancers is co-occupancy of multiple tissue-specific TFs (He et al., 2011; Palstra and Grosveld, 2012). Upstream TFs regulating enhancer activity have been assessed by various methods including using clustered, regularly interspaced, short palindromic repeat (CRISPR)/CRISPR associated protein 9 (CRISPR/Cas9)-based gene editing approaches (Jinek et al., 2012; Cong et al., 2013). CRISPR/Cas9 is typically targeted to a genomic locus using one to four single guide RNAs (sgRNAs) to generate DNA double strand breaks (DSBs) in the vicinity of a genomic region to be deleted, such as an enhancer (Sander and Joung, 2014; Mianne et al., 2017). After generating single DSBs, short deletions within enhancers allowed specific upstream TFBS to be identified (Canver et al., 2015; Sanjana et al., 2016; Kvon et al., 2016; Korkmaz et al., 2016). A more commonly used approach is to assess TFs directly binding at enhancers by chromatin immunoprecipitation (ChIP) methods including ChIP-qPCR (Nishida et al., 2005) and ChIP-seq (Robertson et al., 2007; Jothi et al., 2008). Other techniques for identifying TFs binding to specific DNA sequences include DNaseI footprinting (Brenowitz et al., 1986) and electrophoretic mobility shift assays (EMSAs) (Garner and Revzin, 1986). Other approaches to identify upstream regulators of enhancer activity employ bioinformatic prediction of TFBS using consensus binding site databases and multi-species conservation (Hertz and Stormo, 1999; Tompa et al., 2005; Wingender et al., 1996; Sandelin et al., 2004). The functional role of a TFBS predicted to be important for enhancer function can be assessed using mutated enhancer-reporter constructs (Small et al., 1991; Vincent et al., 2016).

An alternative bioinformatic approach to TFBS identification is DNaseI-seq digital footprinting (Hesselberth et al., 2009; Neph et al., 2012). DNaseI footprinting relies on the simple principle that DNA bound by TFs is protected from nuclease digestion (Galas and Schmitz, 1978). Observing small regions protected from DNaseI digestion within a regulatory element can be used to identify TFBS without prior knowledge of which TFs bind (Galas and Schmitz, 1978; Hesselberth et al., 2009). It also allows TFBS that are differentially bound in different cell types to be identified (Piper et al., 2015). A downside of digital DNaseI footprinting is that it requires extremely high DNaseI-seq data quality and sequencing depth (Schwessinger et al., 2017). To alleviate these shortcomings, meta approaches have been developed that allow genome wide 'pile-up' of DNaseI footprint signals of short sequences to increase signal-to-noise (Pique-Regi et al., 2011; Maurano et al., 2015; Schwessinger et al., 2017).

1.4.3.1 Upstream regulators of haematopoietic *Runx1* enhancers

Multiple well-known haematopoietic TFs bound to *Runx1* enhancers by ChIP-seq in haematopoietic cell lines expressing *Runx1* (Nottingham et al., 2007; Schütte et al., 2016). These TFs included E-box TFs Scl/Tal1 and Lyl1, Ets factors Pu.1, Erg and Fli1, Homeobox protein Meis1, and Zinc-finger TFs Gfi1b and Gata2 (Figure 1.4 c, Nottingham et al. 2007; Schütte et al. 2016). ChIP is a powerful technique for discovering upstream TFs regulating an enhancer but requires prior knowledge about which TFs to target and for ChIP-grade antibodies to be available. Moreover, it requires large numbers of cells making binding of haematopoietic TFs to *Runx1* enhancers in populations undergoing EHT *in vivo* challenging to assess. An alternative

approach to analysis of TF binding to enhancers is to use sequence-based prediction of TFBS (Wasserman and Sandelin, 2004). Through mutation of predicted TFBS, several upstream factors were shown to be important for Runx1 +23, +110, or +204 enhancer function including E-box, Ets, Gata, and Runx motifs (Figure 1.4 d). While analysing sequence conservation improves the accuracy of functional TFBS prediction, not all TFs are represented within motif databases (Wasserman and Sandelin, 2004). Unbiased approaches like digital DNaseI footprinting are yet to be used for the identification of upstream TFBS in *Runx1* enhancers.

Runx1 has been suggested to undergo autoregulation (Nottingham et al., 2007; Schütte et al., 2016; Martinez et al., 2016). Indeed, auto-regulatory loops are pervasive features of transcriptional regulatory networks (Kielbasa and Vingron, 2008), and various developmentally regulated TFs engage in autoregulation such as foxd3 (Lukoseviciute et al., 2018), Gata1 (Huang et al., 2005), and Hox genes (Papadopoulos et al., 2019). Mutation of Runx motifs in the P1 promoter led to decreased Runx1 P1 activity in cell lines, indicating that *Runx1* engages in a positive feedback loop (Martinez et al., 2016). Runx motifs in the 5'UTR, on the other hand, acted to inhibit Runx1 P1 transcription, indicative of Runx1 engaging in a negative feedback loop (Martinez et al., 2016). The ability for Runx1 to recruit polycomb repressive complex 1 (PRC1, Yu et al. 2012) offers a plausible mechanism for Runx1 to engage in auto-repression.

As well as binding to its own promoter, autoregulation of *Runx1* also occurs via binding to its own enhancers (Nottingham et al., 2007; Schütte et al., 2016). Interestingly, Runx motifs seemed to have opposing activities when the same enhancer was tested in different developmental settings. For example, Runx motifs were required for +23 enhancer function in transient transgenic enhancer-reporter cell lines (Nottingham et al., 2007), but not in stable transgenic enhancer-reporter lines (Schütte et al., 2016) and mouse embryos (Nottingham et al., 2007). This suggests that the experimental model used to examine enhancer-reporter assays may be important. A different set of cofactors will be present in different cell types, possibly leading to differences in the outcome of Runx1 binding to an enhancer. Moreover, the fact that Runx1, Runx2, and Runx3 all bind the same canonical motif could allow for different Runx factors to bind Runx motifs in different cell types, possibly also contributing to differences in enhancer activity after Runx motif mutation.

Runx motifs also exhibited opposing roles across different Runx1 enhancers tested in the same cellular context (Schütte et al., 2016). For example, mutating the Runx TFBS in the +23 enhancer increased its activity, while mutating the same TFBS in +204 decreased its activity in stable enhancer-reporter transgenic cell lines (Figure 1.4 d, Schütte et al. 2016). Similarly, foxd3 was previously shown to be capable of either activating or repressing its own enhancers in an enhancer-specific manner (Lukoseviciute et al., 2018). This suggests that differences in co-factor recruitment to enhancers may alter the function of a TF binding across different enhancers even within the same developmental setting.

1.4.4 Super enhancers

It has been suggested that clusters of multiple enhancers may act synergistically to upregulate the expression of developmentally regulated genes. Such clusters of en-

hancers have been termed 'super enhancers' (SEs) (Hnisz et al., 2013; Loven et al., 2013; Whyte et al., 2013) or 'stretch enhancers' (Parker et al., 2013) and are typified by exceptionally high levels of Mediator and lineage-specific TF binding (Loven et al., 2013; Whyte et al., 2013). SEs were first identified as important pluripotency regulators in mESCs (Whyte et al., 2013) and were subsequently identified in other cell types including multiple myeloma (Loven et al., 2013). They were suggested (Kim et al., 2014; Whyte et al., 2013) to be similar to the locus control regions (LCRs) that were previously identified as important regulators of many haematopoietic genes including β-globin (Grosveld et al., 1987; Forrester et al., 1987; Noordermeer and de Laat, 2008).

The SE model proposes that multiple enhancers can act synergistically to upregulate gene expression in a manner beyond what each of the individual enhancers could contribute (Hnisz et al., 2013). Evidence in support of the SE model has come from studying dynamically regulated haematopoietic genes. Deletion of either one of two of the enhancers in the Th2 cytokine gene locus (containing the genes *IFN-*γ, *IL-4/5/13*) LCR led to a drastic reduction in Th2 cytokine gene transcription (Kim et al., 2014; Lee and Flavell, 2005). Additionally, deletion of either one of three immunoglobulin-κ enhancers led to reduced enhancer-promoter interactions and gene transcription in B cells (Jiang et al., 2016; Proudhon et al., 2016). SEs have also been described in a non-haematopoietic context, such as in the broadly expressed *Fgf8* gene (Marinic and Spitz, 2013), mammary gland (Shin et al., 2016), hypothalamic neurons (Lam et al., 2015), dorsal root ganglia (Appel et al., 2016), and mESCs (Hnisz et al., 2015). Mechanistically, deletion of a single enhancer in a SE led to loss of active marks at the neighbouring enhancers (Hnisz et al., 2015). Together these data suggest that under certain circumstances an interdependency exists between clusters of enhancers. It is unclear under what circumstances this principle of synergistic enhancer function does and does not apply to.

A key study specifically tested the hypothesis that a cluster of enhancers acts synergistically to regulate α-globin expression in erythroid cells (Hay et al., 2016). Erythroid-specific expression of α-globin is regulated by a cluster of enhancers that was classified as a SE in erythroid cells based on exceptionally high Med1 occupancy (Hay et al., 2016). However, when a single enhancer of the α-globin enhancer cluster was deleted in erythroid cells *in vivo*, gene expression was impacted only modestly (Anguita et al., 2002; Hay et al., 2016). However, when two α-globin enhancers were deleted, a more dramatic reduction in transcription was seen (Hay et al., 2016). The reduction in α-globin transcription observed after dual enhancer deletion was the same as the combination of the individual enhancer deletion effects, showing that the enhancers are acting additively and not synergistically. Therefore, despite meeting the computational criteria of a SE, the α-globin enhancers do not functionally behave as one *in vivo* (Hay et al., 2016).

Further evidence at other genes also supports that enhancers may not always act synergistically. The β-globin LCR enhancers were previously also suggested to be acting additively and not synergistically (Bender et al., 2012). Early work in T cells revealed that single enhancer deletions in the TCR-γ locus produced only small reductions in gene expression (Xiong and Raulet, 2002). More recently, a *Myc* SE cluster specific

to blood progenitor cells was shown to harbour multiple enhancer elements each contributing additively to overall *Myc* expression levels (Bahr et al., 2018). These studies at haematopoietic gene loci seemingly argue against the SE model. Evidence against the SE model has also come from different developmental contexts including the developing limb bud. A recent study suggested that most enhancers controlling limb bud development are behaving additively and not synergistically, with each additional enhancer deletion producing a linear reduction in transcriptional output (Osterwalder et al., 2018). It has been suggested that the criteria used to identify SEs might simply identify strong enhancers of genes important for instructing cell fate but might not define a functionally distinct class of regulatory elements (Hay et al., 2016; Pott and Lieb, 2015).

1.4.4.1 *Runx1* Super Enhancer

Runx1 enhancers contained within intron 1 (containing +3, +23, +110 enhancers) or within the gene desert adjacent to the gene have previously been classified as being part of a SE by meeting bioinformatic criteria such as high H3K27ac enrichment (Mill et al., 2019; Gunnell et al., 2016; Schuijers et al., 2018; Hnisz et al., 2017; Saint-André et al., 2016; Kwiatkowski et al., 2014). However, only one study tested whether these enhancers act synergistically or additively to regulate *RUNX1* expression (Mill et al., 2019). Deletion of the +23 enhancer in a human leukaemia cell line reduced *RUNX1* expression. This suggests that *RUNX1* enhancers may be acting synergistically to drive expression. Further work is required to clarify whether this is the case in contexts other than the leukaemia cell line used (Mill et al., 2019). At many loci, like *Runx1*, many different enhancers may be present and deletion of multiple of them is sometimes required for gene transcription to be significantly impacted (Hay et al., 2016). Could these enhancers be acting redundantly?

1.4.5 Shadow enhancers

The term "shadow enhancer" was introduced to describe functionally redundant enhancers of *Dorsal* (Hong et al., 2008), *Shavenbaby* (Frankel et al., 2010), and *Snail* (Perry et al., 2010) in Drosophila. In the shadow enhancer model, several redundant enhancers each perform similar regulatory functions for a gene, allowing loss or mutation of one enhancer to be compensated for by another (Hong et al., 2008; Frankel et al., 2010; Perry et al., 2010). This has been shown to provide phenotypic robustness (Frankel et al., 2010; Perry et al., 2010) and hypothesised to allow the evolution of new regulatory functions (Hong et al., 2008).

Functionally redundant enhancers active during development have been identified in mammals by performing enhancer deletions in the native chromatin environment and observing only mild defects in gene transcription levels. Functional enhancer redundancy has been reported for several genes including β-globin (Bender et al., 2012), *Hoxd* (Montavon et al., 2011), *Shh* (Yao et al., 2016), and *Runx3* (Appel et al., 2016). However, there are also examples where a single enhancer is absolutely required to provide important gene regulatory functions, such as the *Gata2* +9.5 enhancer that is critical for *Gata2* expression during HSC ontogeny (Gao et al., 2013). Developmentally regulated transcription factors have been suggested to be regulated

by more shadow enhancers compared to other genes (Cannavò et al., 2016). However, the interpretation of functional enhancer redundancy is likely to be sensitive to the exact cell type and developmental stage at which it is assayed.

As well as through performing enhancer deletions, shadow enhancers have previously been identified when enhancers produced overlapping expression patterns in ectopic enhancer-reporter studies. For example, hundreds of *cis*-regulatory elements that were predicted to be enhancers in brain (based on histone modifications) across different neuron-specific gene loci exhibited similar and partially overlapping enhancer-reporter expression patterns (Visel et al., 2013).

1.4.5.1 Possible shadow enhancers of *Runx1*

A study involving our laboratory assayed transient enhancer-reporter embryos for predicted enhancers of nine haematopoietic TFs, including *Runx1* (Schütte et al., 2016). On average, each TF had at least three *cis*-regulatory elements and some, but not all of these produced partially redundant and overlapping expression patterns (Schütte et al., 2016). As detailed in section 1.4.2.1, in the case of *Runx1* haematopoietic enhancers, -59, +3, +23 exhibited partially overlapping activity during EHT (Schütte et al., 2016). While two other *Runx1* enhancers, +110 and +204, had distinct expression patterns (unpublished observations, de Bruijn laboratory). It is important to note that an overlapping enhancer-reporter expression pattern does not necessarily equate to redundancy at the functional level (Marinic and Spitz, 2013). Therefore, it will be important to perform genomic enhancer deletions of haematopoietic *Runx1* enhancers *in situ* to investigate functional redundancies between them during EHT.

1.4.6 Assessing functional requirements of enhancers

Classically, it was understood by analysing naturally occurring mutations in people with β-thalassemia that deletion of an enhancer in its endogenous locus can lead to loss of gene expression (Kioussis et al., 1983; Driscoll et al., 1989; Weatherall, 2001; Li et al., 2002; Higgs et al., 2012). In the laboratory setting, enhancers have been scrutinised through targeted perturbations of enhancers at endogenous loci. Early studies employed homologous recombination (HR) in mESCs for this, and convincingly showed that evolutionarily conserved enhancers are sometimes functionally required for developmentally regulated gene expression (for example, Reik et al. 1998; Xiong and Raulet 2002; Anguita et al. 2002; Sagai et al. 2005). More recently, clustered, regularly interspaced, short palindromic repeat (CRISPR)/CRISPR associated protein 9 (CRISPR/Cas9)-based gene editing approaches (Jinek et al., 2012; Cong et al., 2013) have been adopted to functionally interrogate enhancers of many different genes including *Sox2* (Li et al., 2014b), *Gata2* (Huang et al., 2016), *Myc* (Bahr et al., 2018) and multiple genes expressed in developing mouse limbs (Osterwalder et al., 2018). CRISPR/Cas9 has been widely used for performing enhancer deletions, but is not without its caveats.

1.4.7 Considerations for the use of CRISPR/Cas9-based genome editing

To reduce mutations occurring when CRISPR/Cas9-induced DSBs are generated off-target, modified approaches have been developed including CRISPR/Cas9^{D10A} nickase (Ran et al., 2013). DSBs can be repaired through several endogenous repair pathways including HR, non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ), or single strand annealing (SSA) (Sander and Joung, 2014; Symington and Gautier, 2011; McVey and Lee, 2008). If a repair template is provided, targeted edits including enhancer deletions or insertions are achievable through homologous recombination (Kvon et al., 2016). Without an exogenous repair template, DNA repair of two DSBs flanking an enhancer often results in loss of the intervening sequence, leading to deletion of the entire enhancer sequence. Such deletions are commonly detected by short-range PCR (SR-PCR, with primers that are typically $\leq$ 200 bp away from sgRNA target sites) and Sanger sequencing (Mianne et al., 2017; Hendel et al., 2015). Insertions or deletions (indels) can also result after DSB repair and are typically small ($<$ 50 bp) (Cradick et al., 2013; Lieber, 2010; Koike-Yusa et al., 2014; van Overbeek et al., 2016; Tan et al., 2015). Larger deletions (LDs), however, are also sometimes seen (Cradick et al., 2013; Wang et al., 2013; Canver et al., 2014; Ma et al., 2014; Zhou et al., 2014; Parikh et al., 2015; Zhang et al., 2015; Shin et al., 2017; Birling et al., 2017; Codner et al., 2018; Kosicki et al., 2018; Adikusuma et al., 2018; Chakrabarti et al., 2019; Owens et al., 2019).

Rather than deleting an enhancer in its entirety using Cas9 nuclease, alternative approaches have also been used to interrogate enhancer function. Catalytically dead Cas9 (dCas9) (Qi et al., 2013) fused to transcriptional repressors KRAB (Thakore et al., 2015) or LSD1 (Kearns et al., 2015) was used to identify functional enhancers for genes including *GATA1* and *MYC* (Fulco et al., 2016), *Oct4* (Kearns et al., 2015), and *β-globin* (Thakore et al., 2015), among others. Care should be taken when examining enhancer redundancy in the laboratory through perturbations such as CRISPR/Cas9 genome editing, however, because an enhancer that is redundant under experimental conditions might be required under more stressful 'real world' conditions.

1.5 How do enhancers function to regulate gene transcription?

The mechanisms by which enhancers functionally activate distant gene promoters has been a topic of much debate (Plank and Dean, 2014). A key study in erythroid cells provided important evidence for how enhancers can function in *cis* at a significant genomic distance (Tolhuis et al., 2002). Erythroid-specific β-globin gene expression is controlled by a cluster of enhancers (LCR) located over 50 kb away from the gene promoter (Forrester et al., 1987; Grosveld et al., 1987). Using chromosome conformation capture (3C) (Dekker et al., 2002), β-globin LCR enhancers were detected physically interacting with the β-globin promoter despite the genomic distance between them (Tolhuis et al., 2002). 3C utilises proximity-ligation of restriction enzyme-digested chromatin fragments to infer spatial proximities between *cis*-regulatory elements in living cells (Dekker et al., 2002). Depending on the methodology used, frequencies of ligated restriction fragments can be detected using PCR or direct sequencing. Build-

ing on the basic principles of 3C, a plethora of techniques for dissecting chromosome conformation have been developed including but not limited to 4C (Simonis et al., 2006; Zhao et al., 2006), 5C (Dostie et al., 2006), ChIA-PET (Fullwood et al., 2009), Hi-C (Lieberman-Aiden et al., 2009; Dixon et al., 2012), Single-cell Hi-C (Nagano et al., 2013), Capture-Hi-C (Dryden et al., 2014), Capture-C (Hughes et al., 2014), Next Generation (NG) Capture-C (Davies et al., 2015), Micro-C (Hsieh et al., 2015), HiChip (Mumbach et al., 2016), Tri-C (Oudelaar et al., 2018), and Tiled-C (Oudelaar et al., 2019).

Importantly, interactions in *cis* between the β-globin promoters and LCR that were detected in erythroid cells were not seen in non-erythroid cell types that did not express the gene (Tolhuis et al., 2002). The cell-type specificity of enhancer-promoter interactions implied that these interactions were functionally important for gene transcription. Indeed, enhancer-promoter interactions are not merely correlated with gene activation but appear to be causative. Forced interactions between the LCR and β-globin promoters was sufficient to drive expression even in cells where this would not normally occur (Deng et al., 2012). This suggests a model in which enhancer-promoter interactions directly increase promoter transcriptional output. Further evidence in support of this has since been observed at a variety of loci in different contexts, including developing limbs (Amano et al., 2009; Montavon et al., 2011; Lupianez et al., 2015), lung (Tsukiji et al., 2014), and brain (Giorgio et al., 2015). Strikingly, however, recent evidence suggests that an enhancer can upregulate gene expression without physically contacting a promoter it regulates (Benabdallah et al., 2019). On the whole, however, enhancers seem to act on promoters by physically contacting them in 3D space.

How an enhancer physically contacting a promoter leads to transcriptional upregulation is incompletely understood. Enhancers are thought to promote transcription by acting at distinct stages of the highly regulated transcription cycle (Fuda et al., 2009; Plank and Dean, 2014). An early study of the β-globin LCR suggested that enhancers may release paused Pol II at promoters, promoting the transition from transcription initiation to elongation (Sawado et al., 2003), a finding that has since been observed for other enhancers (Liu and Rosenfeld, 2013; Chen et al., 2017). Other work suggests that enhancers might act to promote transcription initiation through interaction with the basal transcriptional machinery (Koch et al., 2011; Liu et al., 2011; Ren et al., 2011; Plank and Dean, 2014). A recent study in erythroid cells found that GATA1-dependent enhancers act to remove the polycomb group protein repressive histone mark H3K27me3 from gene promoters and thus maintain transcription (Saxena et al., 2017). Together these data suggest that enhancers can influence gene transcription through diverse mechanisms.

However, enhancer-promoter interactions have not always been reported to be functional. In developing limb buds, low level *HoxD* enhancer-promoter contacts were detected in cell types that did not express *HoxD* genes (Andrey et al., 2013; de Laat and Duboule, 2013; Montavon et al., 2011). This chromatin organisation likely reflects non-specific and diffuse DNA interactions between a gene promoter and its local chromatin neighbourhood. All DNA "has to be somewhere", and given the requirement for human nuclei to contain $\sim$2 meters of linear DNA molecules (Fraser et al.,

2015), local non-specific interactions occurring even when transcription is inactive are inevitable. Given that local diffuse interactions are happening all the time throughout the genome, what delimits the range and specificity of *cis*-interactions between promoters and enhancers is poorly understood.

1.6 Structural considerations of gene expression

Breakthrough studies using Hi-C, an all-vs-all 3C technique (Lieberman-Aiden et al., 2009), revealed that the eukaryotic genome is compartmentalised at the scale of 100 kb to 1 Mb size topologically associating domains (TADs) (Dixon et al., 2012, 2015; Lieberman-Aiden et al., 2009; Nora et al., 2012; Rao et al., 2014; Vietri Rudan et al., 2015). Sequences within a TAD interact more frequently within the TAD than with sequences outside of it and are also called self-interacting domains (Szabo et al., 2019). Enhancer-promoter interactions are generally thought to occur within TADs (Lettice et al., 2011; Symmons et al., 2014; Lupianez et al., 2015). Indeed, naturally occurring TAD boundary deletions are associated with pathogenic "enhancer adoption"—causing limb and brain developmental abnormalities in humans (Lettice et al., 2011; Lupianez et al., 2015; Franke et al., 2016). TADs were originally considered the basic building blocks of the eukaryotic genome because they appeared generally invariable between different cell types and largely conserved during evolution (Dixon et al., 2015; Lieberman-Aiden et al., 2009; Rao et al., 2014). How TADs are formed and maintained inside living cells are both areas of great interest.

1.6.1 Loop extrusion model of chromatin loop formation

Recent work has focussed on the role of structural DNA binding proteins in 3D genome organisation. In particular, CCCTC-Binding Factor (CTCF) draws much attention as an important factor involved in chromatin organisation (Phillips and Corces, 2009). CTCF was first identified as a transcriptional repressor of *Myc* (Lobanenkov et al., 1986, 1990), an oncogene associated with Burkitt's lymphoma (Finver et al., 1988; Molyneux et al., 2012). It was shown later that CTCF binds mammalian insulator elements and is required for their function (Bell and Felsenfeld, 1999). Insulator elements block enhancer activity when placed between an enhancer and a promoter driving a reporter gene in plasmid-based insulator assays (Bell and Felsenfeld, 1999; Chung et al., 1993, 1997). The insulating properties of CTCF have also been observed in endogenous chromatin as CTCF was found to be strongly enriched at TAD boundaries (Dixon et al., 2012; Rao et al., 2014).

How CTCF is able to constrain genomic interactions is an area of intense research. CTCF proteins homomultimerised *in vivo* (Yusufzai et al., 2004), suggesting that distant CTCF sites interacting with one another could play a role in this insulation. The loop extrusion model of 3D chromatin loop formation is currently the most important model for how distant CTCF sites are brought into proximity (Figure 1.5 a) (Nasmyth, 2001; Alipour and Marko, 2012; Sanborn et al., 2015; Fudenberg et al., 2016). In the loop extrusion model, a loop-extruding factor (LEF) is loaded onto DNA and then actively translocates along the DNA fibre until reaching a barrier, potentially a CTCF site, where translocation stalls. One candidate LEF is cohesin, a member of the structural maintenance of chromosome (SMC) protein family that

is required for sister chromatid cohesion during metaphase (Nasmyth and Haering, 2005). During interphase, cohesin is found at many of the same loci as CTCF genome-wide in a variety of cell types (Parelho et al., 2008; Rubio et al., 2008; Stedman et al., 2008; Wendt et al., 2008), and was recently visualised actively extruding loops of DNA (Davidson et al., 2019). In support of cohesin's role as a LEF, its depletion by short interfering RNA (siRNA) or the auxin-inducible degron system reduced CTCF-mediated long-distance chromatin interactions (Hadjur and Merkenschlager, 2009; Nativio et al., 2009; Wendt et al., 2008) and TAD boundary strength (Rao et al., 2017; Wutz et al., 2017; Schwarzer et al., 2017). Once cohesin has extruded a chromatin loop, translocation of DNA is thought to be stalled when encountering an appropriately oriented CTCF-bound anchor site (Rao et al., 2014; de Wit et al., 2015; Guo et al., 2015; Vietri Rudan et al., 2015; Fudenberg et al., 2016). In support of this, 'stripes' of interaction have been observed emanating from CTCF-bound anchor sites in Hi-C data (Vian et al., 2018). Taken together, mounting evidence points to SMC complexes such as cohesin functioning as LEFs to colocalise distant CTCF sites in a manner consistent with the loop extrusion model.

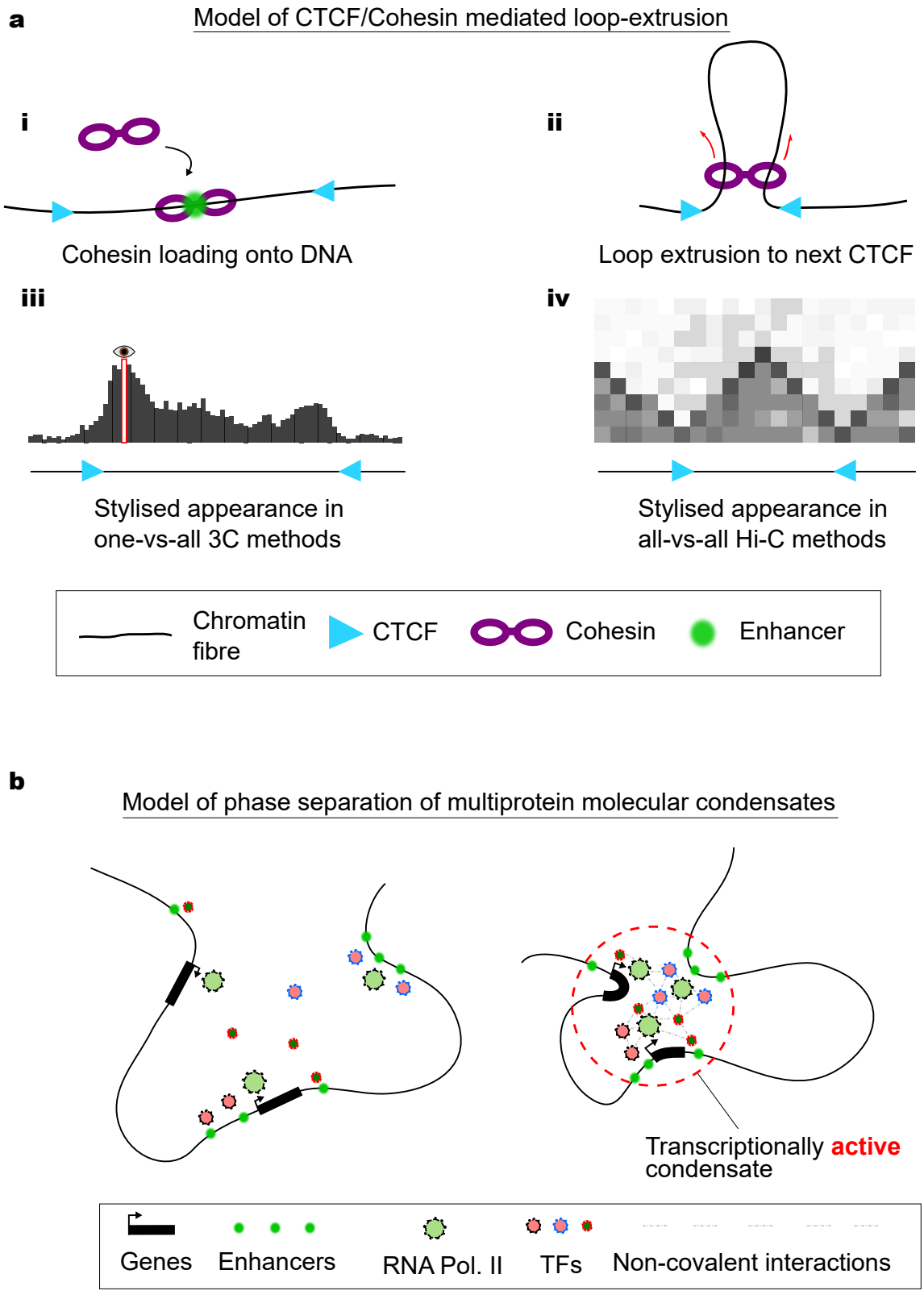


Figure 1.5 – Chromatin organisation by loop extrusion and phase separation. *Legend continued on next page.*

Figure 1.5 – *Legend continued from previous page.* a) (i) The loop extrusion model of chromatin loop formation. In the model, cohesin (depicted as a purple double ring)is loaded onto the chromatin fibre, possibly aided by active enhancers (shown as a green circle). (ii) Cohesin then translocates the chromatin fibre to extrude a loop of DNA until reaching an appropriately oriented CTCF binding site (shown as a blue arrowhead). Due to the dynamic nature of cohesin being loaded and unloaded often, the frequency of interactions within the chromatin loop will be increased, creating a self interacting domain. Stylised representations of self-interacting domains bounded by convergently oriented CTCF sites are shown in the style of one-vs-all (iii, viewpoint indicated by a brown eye icon) and all-vs-all (iv) chromosome conformation capture-based techniques. b) Phase separation model of transcriptional control. Multiple genes (thick black lines with arrows representing promoters) and enhancers (green circles) are distributed throughout a region of chromatin (thin black line). Different transcriptional activators including tissue specfic transcription factors (TFs, small coloured circles as indicated) and RNA polymerase II (RNA Pol. II, larger green circles surrounded by black lines) are indicated. Each component engages in multiple low affinity non-covalent interactions (grey dashed lines between activators) that overall form a transcriptionally active condensate. The formation of the condensate facilitates distant regions along the chromatin fibre to be brought into proximity (compare the relative distances between genes and enhancers in the left diagram to the right diagram). In this example the condensate contains genes and enhancers located along a contiguous chromosome, but the same mechanism would hold true for interactions between chromosomes.

Other proteins aside from CTCF and cohesin have also been implicated in fostering chromatin interactions in *cis*. Other factors including YY1 (Beagan et al., 2017; Weintraub et al., 2017), and Mediator (Allen and Taatjes, 2015; Jeronimo et al., 2016; Kagey et al., 2010; Petrenko and Struhl, 2016) have been implicated in bridging genomic distances. Moreover, loop extrusion does not explain all of the structures observed in the nucleus. Hi-C data revealed that distinct chromosomal compartments exist that are made up of several TADs that appear to share functional features (Lieberman-Aiden et al., 2009). Compartments can be segregated into either transcriptionally active (A) or inactive (B), (Lieberman-Aiden et al., 2009; Bickmore and van Steensel, 2013; Bonev and Cavalli, 2016), that generally correlate with early (A) or late (B) DNA replication timing (Ryba et al., 2010; Sima et al., 2019), and are generally either located at the nuclear interior (A), or associated with the nuclear lamina (Guelen et al., 2008; Dixon et al., 2012; Rao et al., 2014; Fraser et al., 2015). Intriguingly, acute loss of CTCF/cohesin does not impact this higher order compartmentalisation of the genome (Nora et al., 2017; Rao et al., 2017; Wutz et al., 2017; Schwarzer et al., 2017; Sima et al., 2019). This suggests that mechanisms other than loop-extrusion are likely to be involved in the establishing the higher order structures like compartments observed within the nucleus.

1.6.2 Phase separation model of chromatin architecture

Phase separation is the physical process whereby a single phase substance (eg a liquid) passively separates into at least two distinct phases (Erdel and Rippe, 2018). A classic example of a phase separated organelle within the nucleus is the nucleolus (Feric et al., 2016). Recently, the principles of phase separation have been applied to understanding mechanisms of transcriptional control (Hnisz et al., 2017). Phase separation invokes the notion of multivalency, which posits that large stable complexes could form from the combined forces of multiple smaller components each with relatively weak binding affinities (Ruthenburg et al., 2007). It is thought that specific sequences and/or histone modifications at enhancers and/or promoters act as nucleation sites to initiate binding of transcriptional activators (Figure 1.5 b) (Hnisz et al., 2017; Sabari et al., 2018; Cho et al., 2018; Boehning et al., 2018; Chong et al., 2018). In the model, given a critical threshold concentration of these molecular components, multivalent electrostatic protein-protein interactions between the individual components facilitates phase separation into a discrete liquid droplet (Erdel and Rippe, 2018); one phase forming with a high concentration of the molecular components (transcriptionally active condensate, Figure 1.5 b), while the other is depleted of these components (Hnisz et al., 2017; Erdel and Rippe, 2018).

Phase separation offers an explanation for several intriguing aspects of chromatin biology. Firstly, it agrees with a much earlier observation that multiple discrete foci (termed 'transcription factories') containing twenty or more active Pol II complexes were seen in fixed HeLa cells (Iborra et al., 1996). Phase separation also helps to understand why compartments may still be present in Hi-C data even without TADs after cohesin or CTCF loss (Nora et al., 2017; Rao et al., 2017; Wutz et al., 2017; Schwarzer et al., 2017) because large multivalent protein-protein interactions will be stable over time. The phase separation model may also explain how long-range interactions that are incompatible with loop extrusion may occur, such as *trans* inter-

actions (Spilianakis et al., 2005; Simonis et al., 2006; Brown et al., 2006; Lomvardas et al., 2006; Link et al., 2013; Kim et al., 2014; Lim et al., 2018; Monahan et al., 2019) or long-range *cis*-interactions over multiple Mb (Simonis et al., 2006; Proudhon et al., 2016). Other observations, such as the observed insulating functions of CTCF binding sites located within TADs (de Wit et al., 2015; Hansen et al., 2017) are less well understood in the context of phase separation. As such, the exact molecular drivers behind phase separation and its significance for transcriptional regulation are still being elucidated.

1.6.3 How important is genomic structure for facilitating gene regulation?

Alongside their roles in maintaining chromatin structure, CTCF and cohesin are both implicated in transcriptional regulation (Zheng and Xie, 2019; Schoenfelder and Fraser, 2019). However, important recent work showed that acute and near-complete genome-wide loss of CTCF and cohesin also had relatively mild impacts on global transcription in cell lines (Nora et al., 2017; Rao et al., 2017; Wutz et al., 2017; Schwarzer et al., 2017; Hyle et al., 2019). Interestingly, 23 of the 49 genes that were downregulated after cohesin depletion were within 500 kb of an annotated SE (Rao et al., 2017), suggesting that cohesin may be important for facilitating a subset of enhancer-promoter interactions. After genome-wide depletion of CTCF for 24 hours in mESCs, only ∼185 genes were downregulated (Nora et al., 2017). Out of these downregulated genes, 80% were bound by CTCF within 1 kb of the TSS (Nora et al., 2017), suggesting a direct role for CTCF in regulating these genes. However, indirect and direct effects are difficult to disentangle in these genome-wide depletion studies.

Targeted deletion of specific regions have also been used to elucidate transcriptional regulatory effects of CTCF and cohesin. For example, targeted removal of CTCF binding sites has been shown to reduce the expression of several genes (Guo et al., 2012, 2015; Lee and Dean, 2017; Ren et al., 2017; Schuijers et al., 2018; Vian et al., 2018; Canzio et al., 2019), indicating that specific CTCF sites can play a role in transcriptional regulation. The mechanisms by which CTCF/cohesin elicit their transcriptional regulatory effects are incompletely understood. In one scenario, CTCF binding directly to enhancers and/or promoters might facilitate enhancer-promoter interactions by direct loop formation (Guo et al., 2012, 2015; Lee and Dean, 2017; Schuijers et al., 2018; Canzio et al., 2019; Hyle et al., 2019). However, CTCF binding at promoters was reported to be generally cell type invariant (Schuijers et al., 2018), making tissue-specific gene regulatory effects of CTCF potentially difficult to explain. However, instead of CTCF binding sites being constitutive, it has also been suggested that they might be differentially bound in different cell types, facilitating tissue-specific enhancer-promoter loop formation (Barrington et al., 2019). Methylation of CTCF binding sites has been suggested to antagonise CTCF binding (Bell and Felsenfeld, 2000; Shukla et al., 2011; Wang et al., 2012; Flavahan et al., 2016; Hashimoto et al., 2017; Xu and Corces, 2018; Schuijers et al., 2018; Canzio et al., 2019), providing a possible mechanism for establishing tissue-specific CTCF binding and transcriptional control. Even in the context of constitutive CTCF binding sites, tissue-specific effects could still be seen if CTCF interacts with tissue-specific TFs (Arzate-Mejia et al., 2018), which has previously been observed by co-immunoprecipitation (Delgado-

Olguin et al., 2011; Donohoe et al., 2009; Lee and Dean, 2017). Recent studies at globin genes in erythroid cells support a different model where CTCF sites—rather than facilitating enhancer-promoter interactions directly—impact gene expression by delimiting regulatory compartments (de Wit et al., 2015; Hanssen et al., 2017). Both studies revealed that CTCF sites contribute to the specificity of enhancer-promoter interactions *in vivo*.

Mechanistically, how is it that changing the boundaries of a self-interacting domain through CTCF site deletion only sometimes produces alterations in gene expression? Interestingly, 85% of CTCF-bound sites occur within TADs and only 15% occur at TAD boundaries (Dixon et al., 2012; Rao et al., 2014; Ruiz-Velasco et al., 2017). Moreover, most methylation sites are located within TADs, possibly implying that different CTCF sites may be performing distinct regulatory functions (Van Bortle et al., 2014; Wang et al., 2012; Maurano et al., 2015). However, it has been noted that genome folding occurs in a fractal organisation (Bancaud and Ellenberg, 2012; Gibcus and Dekker, 2013; Oudelaar and Higgs, 2017). As such, intra-TAD CTCF sites may be performing similar boundary functions as TAD boundary CTCF sites, but for smaller scale sub-TADs, which have been characterised through higher resolution analysis of 5C and Hi-C data (Phillips-Cremins et al., 2013; Filippova and Kingsford, 2014; Ramirez et al., 2018). Further, whether a self-interacting domain is defined as a TAD or sub-TAD could simply reflect the resolution of the 3C-type assay or the parameters of a particular TAD calling algorithm, and may not be biologically relevant (Forcato et al., 2017; Oudelaar and Higgs, 2017). Future studies should seek to establish a predictive framework for establishing which CTCF sites may be important to gene transcription.

1.6.4 What came first — structure or transcription?

There is increasing evidence that genomic structures form independently of transcription (Palstra et al., 2008; Wijchers et al., 2016; Hug et al., 2017; Brown et al., 2018). However, changes in transcription could also play a causal role in altering the structure of the genome (Dixon et al., 2012; Hou et al., 2012; Rao et al., 2014; Van Bortle et al., 2014; Moore et al., 2015; Oh et al., 2015; Hug et al., 2017; Ramirez et al., 2018). As well as binding to promoters during gene transcription, Pol II binds at enhancers (Plank and Dean, 2014) and some evidence implicates enhancer-bound Pol II in facilitating enhancer-promoter interactions. Pioneering studies showed that active enhancers produce bi-directionally transcribed non-coding RNA, termed enhancer RNAs (eRNAs) (Kim et al., 2010; Hah et al., 2011; Wang et al., 2011) that are sometimes also spliced (Kowalczyk et al., 2012). It is possible that eRNA transcription may play a role in facilitating enhancer-promoter interactions through Pol II or cohesin recruitment (Li et al., 2013; Hnisz et al., 2017; Kojic et al., 2018). Some work suggests that non-coding transcription could antagonise methylation-dependent CTCF binding sites (Isoda et al., 2017; Canzio et al., 2019), providing a tantalising possible link between the transcription and loop extrusion machinery.

The weight of evidence suggests that structural proteins including CTCF and cohesin are important for facilitating genome interactions at multiple length scales. These interactions seem to be required for establishing the specificity of enhancer-promoter

contacts (Hanssen et al., 2017) possibly through a loop extrusion mechanism. However, distinct mechanisms, possibly mediated by phase separation, may also involved in the maintenance of higher order structures within the nucleus (Hnisz et al., 2017). Mounting evidence also seems to suggest that active transcription might directly impact genomic structure. As such, defining the causal relationships between genome structure and function is still an open debate (Oudelaar and Higgs, 2017; Krijger and de Laat, 2017).

1.7 Outline and aims of thesis

This thesis focusses on the *cis*-regulatory mechanisms of the TF *Runx1* during developmental haematopoiesis. Runx1 is a critical player in developmental haematopoiesis, with complex transcriptional regulation. As such it is an interesting factor to study, to contribute to our understanding of the regulation of large complex genes. For example, do they exhibit similar or distinct regulatory mechanisms to smaller genes with fewer enhancers and promoters that the current models of transcriptional regulation are primarily based on. Moreover, an improved understanding of *Runx1* *cis*-regulation is expected to provide future opportunities to manipulate *Runx1* expression *in vitro* for the *de novo* generation of HSCs or in leukaemia.

My first results chapter (chapter 3) explores the differences in *cis*-interactions that can be observed when *Runx1* switches from transcriptionally inactive to active. It is generally understood that the structure of the genome is important for facilitating proper gene regulation. But what if genomic architecture is simply a necessity for compacting large genomes into a small volume? Then the changes in structure that are correlated with transcription may only be a by-product. However, evolution by natural selection (Darwin, 1872) favours pleiotropy, often reusing genes or pathways for multiple purposes (Preston et al., 2011). Therefore, it seems logical to propose that the mechanisms underlying the structures of the genome may also be directly important for its expression. Many of the genes that have provided key insights into mechanisms of transcriptional regulation are small and relatively simple in structure. To examine the changes in *cis*-regulation that occur when a large and complex gene becomes active, high-resolution chromatin conformation capture at the *Runx1* locus will be performed in cell lines expressing low and high levels of the gene.

The second results chapter (chapter 4) focusses on haematopoietic *Runx1* enhancers. Developmentally regulated TF genes like *Runx1* are generally regulated by a large number of enhancers (Ernst et al., 2011; Cannavò et al., 2016; Osterwalder et al., 2018). However, it is unclear whether each of the enhancers previously identified (Nottingham et al., 2007; Schütte et al., 2016) are equally important for *Runx1* transcription during EHT, or whether certain enhancers may play more or less important role across the different stages of EHT. Moreover, it is currently unclear whether *Runx1* enhancers may be acting synergistically as a SE. Enhancers responsible for *Runx1* transcriptional regulation during EHT will be identified by observing dynamic changes in chromatin accessibility during EHT *in vitro*. To examine whether the enhancers might be regulated by different combinations of upstream factors during EHT, upstream signals converging on *Runx1* enhancers will be investigated. Potential enhancer redundancy and synergy will be explored by deleting enhancers in the endogenous *Runx1* locus in mESCs using CRISPR/Cas9. Detailed genotyping strategies will be developed to allow detection of unwanted LDs and accurate genotyping of genome-edited mESC lines. To explore functional requirements of *Runx1* enhancers during EHT, enhancer knock-out lines will be differentiated *in vitro* to haematopoietic cells.

A possible mechanism for regulating *Runx1* alternative promoter choice is explored in the final results chapter (chapter 5). *Runx1* promoter usage is regulated in a spatiotemporal manner during haematopoietic development (Miyoshi et al., 1995;

Ghozi et al., 1996; Telfer and Rothenberg, 2001; Pozner et al., 2007; Sroczynska et al., 2009a; Bee et al., 2009b, 2010) but regulators of *Runx1* promoter choice are poorly understood. Previous studies have implicated CTCF/cohesin in this process (Horsfield et al., 2007; Marsman et al., 2014; Nora et al., 2017), but it is unclear whether this was via direct or indirect effects. The tissue-specificity of CTCF binding will be examined in the *Runx1* locus that might play a role in *Runx1* alternative promoter choice. DNA methylation has previously been implicated in regulating CTCF binding (Bell and Felsenfeld, 2000; Hashimoto et al., 2017) and promoter activity (Curradi et al., 2002). DNA methylation at the *Runx1* promoters will be examined as a possible candidate for regulating alternative *Runx1* promoter usage and CTCF binding.

2. Materials and methods

2.1 Cell culture

All cell centrifugation steps below were carried out at 200 rcf for 5 mins at room temperature (RT) unless otherwise stated. All cells were maintained at 37 °C and 5% CO_2 in a humidified incubator. All culture media were sterilised by passing through a 0.22 μM filter system (Corning). Cell counting was performed using a neubauer haemocytometer and trypan blue stain (0.4%, Gibco, 15250061) or NucleoCounter® NC-3000™ (Chemometec) according to the manufacturer's instructions.

2.1.1 Culture of mESCs

E14-TG2a (male embryonic stem cell line from the mouse 129/Ola background) (Handyside et al., 1989) and E14-TG2a-RV mESCs (stably transfected with a Venus reporter at the 3' end of Runx1 and a hsp68-mCherry-Runx1+23 enhancer-reporter transgene in the *Col1a1* locus; Lucas Greder, unpublished data) were cultured on 0.1% gelatin (Sigma, G1393) in PBS treated 6-well culture dishes. Culture media consisted of GMEM medium (BHK-21, Gibco, 11710035) supplemented with 100 mM non-essential amino acids (Gibco, 11140050), 100 mM sodium pyruvate (Gibco, 11360070), batch-tested 10% FCS (Gibco), 2% Leukemia Inhibitory Factor (LIF, prepared in house from CHO cells at 1000 Uml^{-1}) conditioned medium, 2 mM L-glutamine (Gibco, 25030-24) and 100 μM β-mercaptoethanol (Gibco, 31350010). Cells were passaged using 0.05% trypsin (Gibco, 25300054) to generate a single cell suspension and split 1:5-1:10 every 2-3 days when they reached 80% confluency.

2.1.2 Culture of 416B cells

The 416B myeloid progenitor cell line (Dexter et al., 1979) was cultured in suspension in T25, T75, or T175 flasks (Corning) using Fischer's medium (Gibco, 21475025) supplemented with 20% horse serum (Gibco, 16050130) and 2 mM L-glutamine. Cells were maintained at densities of between 2 - 8×10^5 cells mL^{-1}. Cells were counted at least every 24 hours.

2.1.3 Culture of HPC7 cells

HPC7 cells (Pinto do O et al., 1998a) were cultured in suspension in T25 flasks (Corning) using IMDM basal medium (Gibco, 12440053) supplemented with 10% FCS (Gibco), 0.15 mM monothioglycerol (Sigma, M6145), 100 $ngmL^{-1}$murine SCF SCF (PeproTech, 250-03, 100 $ng\mu L^{-1}$. Cultures were initiated at a starting density of 5×10^5 cells $mL^{-1}$and maintained below 3×10^6 cells mL^{-1}. Every 48 hours cells were centrifuged before replenishing the entire volume of media to ensure proper SCF activity. Cells were counted at least every 48 hours.

2.1.4 Cryopreservation of cell lines

mESC cells were harvested for cryostorage at 60-80°C confluency by trypsinisation (as in Section 2.1.1). 416B and HPC7 cells were harvested at 60-80% of their maximum cell density. Cells were centrifuged, washed once in cold PBS, and resuspended in 90% FCS (Gibco) and 10% DMSO (Sigma, D2438). One cryovial (Corning) was aliquoted per confluent half of a 6-well plate of mESCs, or per 1×10^7 416B and HPC7 cells, and stored at -80°C using a freezing container (Nalgene® Mr. Frosty, Sigma, C1562). After at least 24 hours at -80°C cells were transferred to liquid nitrogen for long term storage.

2.1.5 Haematopoietic differentiation of mESCs

Directed differentiation of mESCs to haematopoetic lineages was performed using a modified protocol adapted from (Sroczynska et al., 2009b; Pearson et al., 2015). Differentiation medium consisted of StemPro-34 (SP34, Gibco, 10639-011), supplemented with 40X defined serum replacement (Gibco, 10639-011), 0.5 mM ascorbic acid (Sigma A4544), 2 mM L-glutamine (Gibco, 25030-24), 0.45 mM monothioglycerol (Sigma, M6145).

At day 0, mESCs were harvested at 60-70% confluency and in exponential growth phase before beginning differentiation. Cells were washed with 2-3ml PBS before being trypsinised with 1mL prewarmed 0.05% Trypsin-EDTA for 3 mins or until cells visibly detached from plate. Trypsin was inactivated using mESC culture medium without LIF and cells counted a neubauer haemocytometer and trypan blue stain (0.4%, Gibco, 15250061) or NucleoCounter® NC-3000™ (Chemometec) according to the manufacturers instructions. Cells were centrifuged for 5 and 7.5 8 $\times10^5$ cells were seeded into a 100 mm non-tissue culture treated dishes (Corning) containing 20 mL SP34 medium with BMP-4 (R&D, 314-BP-010, stock concentration at 100 ngmL^{-1}) at a final concentration of 5 ngmL^{-1}. Cells were incubated for exactly 72 hours before the removal of 4 mL old medium and the addition of 5 mL fresh SP34 medium. Fresh cytokines were added – bFGF (R&D, 233-FB-025, stock concentration at 100 ngmL^{-1}), activin A (PeproTech, 120-14E, stock concentration at 50 ngmL^{-1})–at a final concentration of 5 ngmL^{-1}each. After a further 24 hours of culture, embryoid bodies (EBs) were harvested by centrifugation and were then washed with PBS before trypsinisation with 0.05 % trypsin for 2 mins at 37°C in a water bath. Tubes containing EBs were agitated every 30 s to prevent them from settling to the bottom of the tube. Trypsinised cells were pipetted several times to ensure a single cell suspension before trypsin was inactivated using mESC culture medium without LIF. Cells were centrifuged, resuspended in FACS buffer containing PBS + 10% FCS, and counted. Cells were stained using rat anti-mouse Flk1-APC antibody (list of antibodies and dilutions in Table 2.1).

After sorting, cells were washed with PBS and centrifuged at 200 rcf for 5 mins at 4°C. Flk1+ cells were plated at a density of 6 $\times10^4$ cells per cm^2 in 0.1% gelatin coated tissue-culture treated plates. Flk1+ cells were cultured with SP34 with the addition of SCF (PeproTech, 250-03, stock concentration at 100 ngmL^{-1}) and VEGF (R&D, 293-VE-010, stock concentration at 100 ngmL^{-1}) at a final concentration of 10 ngmL^{-1}each. Transferrin (Sigma, T8158) was also added at a concentration of

180 μg mL^{-1}. The next morning (12-16 hours) after plating, medium was changed to remove any dead cells post sorting. Cells were cultured for a further 48 hours before harvesting.

After 6 or 7 days total of differentiation, a proportion of cells underwent EHT and were partially adherent and partially in suspension. Cells in suspension were transferred in culture media to 15 mL Falcon tubes and treated with collagenase (Sigma, C0130) at a final concentration of 0.1% for 5 mins at 37°C in a water bath, before pipetting several times, and a further 5 mins incubation. Adherent cells were treated with 0.1% collagenase in FACS buffer for 10 mins in an incubator at 37°C followed by vigorous pipetting to ensure the cell monolayer was completely detached, followed by a further 5 mins incubation. Collagenase was neutralised by the addition of 3 times the volume cold FACS buffer. Suspension and adherent cell fractions were centrifuged for 5 mins at 200 rcf at 4°C resuspended in FACS buffer, and counted before further analysis

2.2 Flow cytometry and cell sorting

Single cell suspensions were stained in FACS buffer (PBS + 10% FCS) for 10-30 mins on ice in the dark. Up to 1 $\times 10^6$ cells were stained in 50 μL buffer. Antibodies used are listed in Table 2.1. After staining, antibody was diluted at least two-fold using FACS buffer, cells were centrifuged again, before resuspending in FACS buffer with the addition of Hoechst 33258 diluted 1:1000.

Table 2.1 – Antibodies used for flow cytometry

Antigen	Fluorophore	Dilution	Supplier	Catalogue number
Flk1	APC	1:100	eBioscience	17-5821-81
CD41	PE	1:400	BD Pharmigen	558040
CD45	APC-eflour780	1:200	eBioscience	13-5821-81
Ter119	PE-Cy7	1:200	BD Pharmigen	557853
VE-cadherin	APC	1:200	eBioscience	17-1441-82

FACS sorting was performed using a Fusion 2 Becton Dickinson machine. Instrument voltages were set up using unstained cells and single stained cells or compensation beads (BD). Gating was performed using fluorescence minus one (FMO) controls. into 100% FCS and kept on ice until plating, which was done as soon as possible.

2.3 Colony forming unit (CFU-C) assays

MethoCult™ (Stem Cell Technologies, 04434) aliquots (2.7 mL) were thawed overnight at 4°C and vortexed thoroughly prior to use. Unsorted cells were harvested from mESC haematopoietic differentiation cultures at day 7 according to Section 2.1.5. 6×10^4 cells were resuspended in 300 μL PBS and gently pipette the thawed MethoCult™

and vortexed thoroughly to mix. MethoCult™ was drawn into a 3 mL syringe with 19G needle attached and dispensed once to remove any bubbles. 1 mL MethoCult™ was dispensed into each one of two sterile 35 mm pipette dishes (Greiner Bio-One, 627161). Lids were replaced and dishes were then rotated several times to evenly distribute the culture medium and cells. Dishes were placed into a 150 mm petri dish (Corning, CLS430599) with another 35 mm dish with lid off containing PBS. CFU-C dishes were incubate at 37°C until counting colonies at day 7-10.

2.4 Immunocytochemistry and confocal microscopy

For imaging mESC differentiation cultures, cells were plated into glass bottom 24-well plates (ibidi, 82406) or 8 chamber slides (ibidi, 80841). Supernatant was removed and kept aside for plating suspension cells onto cover slips. Adherent cells were washed twice in PBS and then fixed at RT for 15 mins using 4% paraformaldehyde solution (Sigma, 252549) diluted in PBS. Adherent cells were incubated with permeabilisation and blocking buffer (buffer) for 1 hour (containing 0.01% Triton X-100 in PBS and 10% FCS). Adherent cells were then incubated with 100-500 μL primary antibodies (list of antibodies and dilutions in Table 2.2) diluted in buffer for 60 mins on ice, rocking at 50 rpm. Cells were washed three times with buffer before staining with secondary antibodies diluted 1:400 in buffer.

Table 2.2 – Antibodies used for immunocytochemistry

Antigen	Fluorophore	Dilution	Supplier	Catalogue number
CD41	unconjugated	1:100	BD Pharmigen	553847
CD45	unconjugated	1:100	BD Pharmigen	550539
Runx1/2/3	unconjugated	1:100	Abcam	ab92336
CD31	unconjugated	1:100	R&D Systems	AF3628
anti-Rat	Alexa Fluor 555	1:400	Invitrogen	A-21434
anti-Goat	Alexa Fluor 647	1:400	Invitrogen	A-21447
anti-Rabbit	Alexa Fluor 647	1:400	Invitrogen	A-11008

2.5 Molecular cloning

2.5.1 Restriction digestion of plasmids

Restriction digestion of plasmids was performed by combining the following in a 1.5mL eppendorf tube:

29 μL H_2O
5 μL plasmid [2 μg total]

2 μL Restriction enzyme
4 μL RE buffer
Total: 40 μL

Restriction enzymes were purchased from NEB or Thermo Fisher Scientific. Digestion reactions were incubated at 37 °C for 1 hour before linearised plasmids were size selected using agarose gel electrophoresis and purified using Zymoclean™ Gel DNA Recovery Kit (Zymo, D4001) according to the manufacturer's instructions.

2.5.2 Plasmid ligation

Plasmids were ligated using standard protocols. Briefly, the following were combined in a 0.2 mL tube.

6 μL H_2O
1 μL insert DNA [3:1 insert:vector molar ratio]
1 μL backbone cut vector [25ng]
1 μL T4 DNA ligase buffer (Invitrogen, 15224017)
1 μL T4 DNA ligase (Invitrogen, 15224017)
Total: 10 μL

Ligation reactions were incubated at 16 °C for 1 hour or overnight at 4 °C. 2 μL of each ligation was transformed into *Escherichia coli* (see Section 2.5.4).

2.5.3 Gibson Assembly

Gibson Assembly (GA) (Gibson et al., 2009) was performed using a custom master mix containing:

100 μL 5X ISO buffer (see below)
0.2 μL 10 UμL^{-1} T5 exonuclease (NEB, M0363S)
6.25 μL 2 UμL^{-1} Phusion polymerase (NEB, M0530S)
50 μL 40 UμL^{-1} Taq ligase (NEB, M0208S)
218.55 μL H_2O

Total: 375 μL

5X ISO buffer contained:

3 mL 1 M Tris-HCl pH 7.5 (Sigma, T2319)
300 μL 1 M $MgCl_2$ (Sigma, M1028)
60 μL 100 mM dATP (Invitrogen, 10297018)
60 μL 100 mM dTTP (Invitrogen, 10297018)
60 μL 100 mM dCTP (Invitrogen, 10297018)
60 μL 100 mM dGTP (Invitrogen, 10297018)
300 μL 1 M DTT (Thermo Scientific, P2325)

600 μL of 50 mM NAD (NEB, B9007S)
1.5 g PEG-8000 (Sigma, 1546605)
H_2O up to 6 mL

Total: 6 mL

GA PCR reactions were performed by combining the following in a 0.2 mL tube and thermal cycling using the following protocol:

31 μL Nuclease Free H_2O
10 μL 5X Phusion HF buffer (NEB, M0530S)
1 μL dNTP mix [10 mM] (NEB, N0447S)
2.5 μL FW GA primer [10 μM]
2.5 μL RV GA primer [10 μM]
2.5 μL template DNA [1 ngμL^{-1} for gBlocks and 100 ngμL^{-1} for gDNA]
0.5 μL Phusion polyermase (NEB, M0530S)
Total: 50 μL

Thermal cycling conditions

a. 98 °C 30 s
b. 98 °C 10 s
c. 68 °C to 60 °C ramp rate set at 0.1 °Cs^{-1}
d. 72 °C 1 min
e. repeat steps b-d for a total of 35 cycles
d. 72 °C 5 min
g. 4 °C ∞

GA PCR primers and template DNA used in individual reactions are included in Tables 2.3 and 2.5. PCR products were purified using DNA Clean & Concentrator Kit (Zymo, D4013) according to the manufacturer's instructions. GA cloning reactions were performed by combining the following in a 0.2 mL tube:

7.5 μL GA master mix (see above)
1 μL 25 ng restriction digested plasmid backbone
1.5 μL purified GA PCR product [equimolar DNA ends to backbone]
Total: 10 μL

GA reactions were incubated for 1 hour at 50 °C before 2 μL was immediately transformed into *Escherichia coli* (see Section 2.5.4).

2.5.4 Plasmid isolation and purification

DH10β chemically competent *Escherichia coli* (Invitrogen, 18297010) were transformed according to the manufacturer's instructions. Ampicillin resistant clones were picked from plates and grown over night in 5 mL LB medium. QIAprep Spin Miniprep Kit (Qiagen, 27104) was used to isolate plasmid DNA and correct inserts were confirmed by Sanger sequencing (Source Bioscience). Correct plasmids were then grown

over night in 50 mL LB and purified using Qiagen plus Midi prep kit (Qiagen, 12943).

2.6 Luciferase assays

2.6.1 pGL3 construct generation

Mutations to Transcription factor binding sites (TFBS) in enhancer sequences were generated *in silico* using the following criteria. Two identical occurrences of AP1 motif TGACTCA in +204 enhancer were mutated to GAGCTAC. One instance of a CEBP motif TTGAGCAA in +110 enhancer was mutated to GCGAGCTG. Four occurrences of Klf motifs were mutated in +23 enhancer. GGGGTGGG mutated to TTCTGCAC, CCCCACCC mutated to AGTACGGA, CCCACCCC mutated to GTACGGAG, GGGTGGGG mutated to TCTGAACC. Mutations to CTCF binding sites in putative insulator sequences were as follows: P2 CTCF motif TGCTGCCCCCTGGCGTCCG mutated to TGCTGCTCAGAGGGCTCCG, -468 CTCF motif AATGTCTCCCCCTGGAGGAC was deleted. All new sequences generated were checked using JASPAR to make sure another TFBS was not inadvertently introduced. Mutated sequences were ordered as gBlocks (IDT) Enhancers were cloned into the SalI site and Insulators were cloned into the XhoI site of pGL3-promoter vector (Promega, E1761) using GA (see Section 2.5.3). GA PCR primers (designed using NEBuilder) were used to add overhanging sequences to the enhancer or insulator gBlocks or wildtype gDNA for each construct as described in Table 2.3. Finished constructs generated by a single GA reaction are separated by a solid line. Where multiple inserts were cloned into a single finished construct, individual PCR reactions (see Section 2.5.3) are separated by a dashed line. 2 μL of each GA reaction (see Section 2.5.3) was transformed into *Escherichia coli* as described in Section 2.5.4.

Table 2.3 – Primers used for Gibson Assembly of luciferase plasmids

Finished construct	Primer 5' to 3' sequence	PCR Template
pGL3 -468-CTCF +23 SV40	cgtgctagcccgggcTACAACTGCCAATGCTCTTC	E14 gDNA
pGL3 -468-CTCF +23 SV40	caccagacGTTCCCGAACTTCTCTGG	E14 gDNA
pGL3 -468-CTCF +23 SV40	tcgggaacGTCTGGTGGGGGTGGGAG	E14 gDNA
pGL3 -468-CTCF +23 SV40	atgcagatcgcagatcAGGTGTCAGCAACCCATCAG	E14 gDNA
pGL3 +23 -468-CTCF SV40	cgtgctagcccgggcGTCTGGTGGGGGTGGGAG	E14 gDNA
pGL3 +23 -468-CTCF SV40	gcagttgtaAGGTGTCAGCAACCCATCAG	E14 gDNA
pGL3 +23 -468-CTCF SV40	tgacacctTACAACTGCCAATGCTCTTC	E14 gDNA
pGL3 +23 -468-CTCF SV40	atgcagatcgcagatcGTTCCCGAACTTCTCTGG	E14 gDNA
pGL3 +23 -468-dCTCF SV40	cgtgctagcccgggcGTCTGGTGGGGGTGGGAG	E14 gDNA
pGL3 +23 -468-dCTCF SV40	gcagttgtaAGGTGTCAGCAACCCATCAG	E14 gDNA
pGL3 +23 -468-dCTCF SV40	tgacacctTACAACTGCCAATGCTCTTC	-468-dCTCF_gBlock
pGL3 +23 -468-dCTCF SV40	atgcagatcgcagatcGTTCCCGAACTTCTCTGG	-468-dCTCF_gBlock
pGL3 P2-CTCF +23 SV40	cgtgctagcccgggcAGAGGCTTTGAGAAGAGATGAG	E14 gDNA
pGL3 P2-CTCF +23 SV40	caccagacCCCGCTACTGTCCACATATTG	E14 gDNA
pGL3 P2-CTCF +23 SV40	gtagcgggGTCTGGTGGGGGTGGGAG	E14 gDNA
pGL3 P2-CTCF +23 SV40	tcgggaacGTCTGGTGGGGGTGGGAG	E14 gDNA
pGL3 +23 P2-CTCF SV40	cgtgctagcccgggcGTCTGGTGGGGGTGGGAG	E14 gDNA
pGL3 +23 P2-CTCF SV40	aagcctctAGGTGTCAGCAACCCATCAG	E14 gDNA
pGL3 +23 P2-CTCF SV40	tgacacctAGAGGCTTTGAGAAGAGATGAG	E14 gDNA

Continued

Table 2.3: continued from previous page

Finished construct	Primer 5' to 3' sequence	PCR Template
pGL3 +23 P2-CTCF SV40	atgcagatcgcagatcCCCGCTACTGTCCACATATTG	E14 gDNA
pGL3 +23 P2-dCTCF SV40	cgtgctagcccgggcGTCTGGTGGGGGTGGGAG	E14 gDNA
pGL3 +23 P2-dCTCF SV40	aagcctctAGGTGTCAGCAACCCATCAG	E14 gDNA
pGL3 +23 P2-dCTCF SV40	tgacacctAGAGGCTTTGAGAAGAGATGAG	P2-dCTCF_gBlock
pGL3 +23 P2-dCTCF SV40	atgcagatcgcagatcCCCGCTACTGTCCACATATTG	P2-dCTCF_gBlock
pGL3 HS4-CTCF +23 SV40	cgtgctagcccgggcCTCCCGGAGACCTTACTTAGAG	Chicken gDNA
pGL3 HS4-CTCF +23 SV40	caccagacCATTCAAGGCCAGGCTGG	Chicken gDNA
pGL3 HS4-CTCF +23 SV40	ccttgaatgGTCTGGTGGGGGTGGGAG	E14 gDNA
pGL3 HS4-CTCF +23 SV40	atgcagatcgcagatcAGGTGTCAGCAACCCATCAG	E14 gDNA
pGL3 +23 HS4-CTCF SV40	cgtgctagcccgggcGTCTGGTGGGGGTGGGAG	E14 gDNA
pGL3 +23 HS4-CTCF SV40	tccgggagAGGTGTCAGCAACCCATCAG	E14 gDNA
pGL3 +23 HS4-CTCF SV40	tgacacctCTCCCGGAGACCTTACTTAGAG	Chicken gDNA
pGL3 +23 HS4-CTCF SV40	atgcagatcgcagatcCATTCAAGGCCAGGCTGG	Chicken gDNA
pGL3 -468-CTCF SV40	cgtgctagcccgggcTACAACTGCCAATGCTCTTC	E14 gDNA
pGL3 -468-CTCF SV40	atgcagatcgcagatcGTTCCCGAACTTCTCTGG	E14 gDNA
pGL3 P2-CTCF SV40	cgtgctagcccgggcAGAGGCTTTGAGAAGAGATGAG	E14 gDNA
pGL3 P2-CTCF SV40	atgcagatcgcagatcCCCGCTACTGTCCACATATTG	E14 gDNA
pGL3 +23 SV40	cgtgctagcccgggcGTCTGGTGGGGGTGGGAG	E14 gDNA
pGL3 +23 SV40	atgcagatcgcagatcGTTCCCGAACTTCTCTGG	E14 gDNA
pGL3 +23-wt SV40	aaatcgataaggatccgGTCTGGTGGGGGTGGGAG	E14 gDNA
pGL3 +23-wt SV40	tctcaagggcatcggCCCCCCCCCAGGTGTCAG	E14 gDNA
pGL3 +23-Gata SV40	aaatcgataaggatccgGTCTGGTGGGGGTGGGAG	+23 gBlock Gata
pGL3 +23-Gata SV40	tctcaagggcatcggCCCCCCCCCAGGTGTCAG	+23 gBlock Gata
pGL3 +23-Klf SV40	aaatcgataaggatccgGTCTGGTGTTCTGCACAG	+23 gBlock Klf
pGL3 +23-Klf SV40	tttgagaagAAAAGAGGTAGTGAGGGG	+23 gBlock Klf
pGL3 +23-Klf SV40	acctcttttCTTCTCAAAGAGCCTGGGATGC	E14 gDNA
pGL3 +23-Klf SV40	tctcaagggcatcggCCCCCCCCCAGGTGTCAG	E14 gDNA
pGL3 +110-wt SV40	aaatcgataaggatccgTCCTCAATCATTGCTCCC	E14 gDNA
pGL3 +110-wt SV40	tctcaagggcatcggAAAATGTCCAACTTCAGGG	E14 gDNA
pGL3 +110-Cebp SV40	aaatcgataaggatccgTCCTCAATCATTGCTCCC	+110 gBlock Cebp
pGL3 +110-Cebp SV40	agtggctgCTACAGCGGAGGTGCAGC	+110 gBlock Cebp
pGL3 +110-Cebp SV40	cgctgtagCAGCCACTCTGTCCTAAG	E14 gDNA
pGL3 +110-Cebp SV40	tctcaagggcatcggAAAATGTCCAACTTCAGGG	E14 gDNA
pGL3 +110-Ets SV40	aaatcgataaggatccgTCCTCAATCATTGCTCCC	+110 gBlock Ets
pGL3 +110-Ets SV40	acccgaaaTATCTTTGTCTGTACTCAGGAGAAG	+110 gBlock Ets
pGL3 +110-Ets SV40	gacaaagataTTTCGGGTGTCGTTGTTG	+110 Ets ultramer
pGL3 +110-Ets SV40	tctcaagggcatcggAAAATGTCCAACTTCAGGG	+110 Ets ultramer
pGL3 +204-wt SV40	aaatcgataaggatccgCAGGGCTCTTTCACAACCC	E14 gDNA
pGL3 +204-wt SV40	tctcaagggcatcggACAGTGGGCCATGCAATG	E14 gDNA
pGL3 +204-Ets SV40	aaatcgataaggatccgCAGGGCTCTTTCACAACCC	+204 gBlock Ets
pGL3 +204-Ets SV40	tctcaagggcatcggACAGTGGGCCATGCAATG	+204 gBlock Ets
pGL3 +204-AP1 SV40	aaatcgataaggatccgCAGGGCTCTTTCACAACCC	+204 gBlock AP1
pGL3 +204-AP1 SV40	tctcaagggcatcggACAGTGGGCCATGCAATG	+204 gBlock AP1

2.6.2 Assaying luciferase activity

Dual Luciferase Reporter Assay System (Promega, E1910) was used to measure luciferase activity normalised to Renilla as an endogenous control. 1×10^7 416B cells were counted centrifuged, washed in PBS, centrifuged again, and then resuspended in 180 μL PBS containing 10 μg pGL3 plasmid and 1 μg Renilla plasmid (pRL-TK, Promega, E2241, both purified as in Section 2.5.4). 1×10^7 cells were electroporated using a Gene Pulser cuvette 0.4 cm electrode gap (Bio-Rad, 1652088), and Gene Pulser (Bio-Rad) with Capacitance Extender (Bio-Rad) set to 960 μFD and 220 mV. After electroporating, 200 μL 416B culture medium was added to the cuvette and the entire volume transferred to a T75 flask (Corning, CLS3276) containing 14 mL warm 416B culture medium. Cells were incubated for 24 hours and then collected in 15 mL tubes (Falcon, 10136120), centrifuged, washed once in PBS, centrifuged again, and supernatant discarded completely. Cells were resuspended in 100 μL 1X passive lysis buffer (Promega, E1910) and incubated for 15 mins at RT in the dark shaking at 50 rpm. Each cell sample was pipetted into three technical replicate wells at 30 μL per well of a 96-well white microwell plate (Thermo Scientific, 236108). 100 μL per well LARII and Stop&Glo buffers were prepared according to the manufacturer's instructions. Luciferase activity was measured using a Fluostar OPTIMA (BMG labtech) with the following settings. Emission filter: Lens, Gain: 3500, Positioning delay: 0.2, No of kinetic windows: 1, Measurement start time: 0, No of interval: 48, Measurement interval time: 0.5, Interval time: 0.5, Volume 1 and 2 set to 100 μL, Pump 1 and 2 set to 310 μL second^{-1}. Fluostar OPTIMA software (BMG labtech) and Excel (Microsoft) was used to analyse data using the following settings. Calc. Range Start 1: 2, Stop 1: 26, Calc. Range Start 2: 28, Stop 2: 48, Range 1 = Table 1,Range 2 = Table 2. Technical triplicate wells were averaged in each experiment and each construct was tested in three separate experiments on different days (n=3).

2.7 CRISPR/Cas9 genome editing

2.7.1 Construct design and generation

Dual-Cas9^{D10A} nickase and Cas9 nuclease-based knock-out strategies were designed using the Zhang lab online tool (crispr.mit.edu) using a previously described approach (Ran et al., 2013). Single guide RNAs (sgRNAs, Table 2.4) contained a G at the 5' position (or the first base was changed to a G) and were ordered as two single stranded oligonucleotides (IDT). Oligos included overhanging sequences (5'-CACCGN...N-3' and 3'-CN...NCAAA-5') to facilitate cloning into BbsI-digested plasmids. Oligos were annealed by combining 8 μL H_2O, 1 μL forward strand oligo, 1 μL reverse strand oligo, and heating to 95 °C for 5 mins, then cooling to 25 °C (ramp rate 0.1 °Cs^{-}1). 1 μL annealed oligo reactions were ligated into BbsI-digested pX335-Neo-GFP (Logan et al., 2015) or pX459 (Ran et al., 2013) plasmids as described in Section 2.5. GA was used to create tandem constructs that contained two or four sgRNAs in one plasmid as described in Section 2.5.3. Primers are described in Table 2.5. Correct sgRNA inserts were confirmed by Sanger sequencing.

Table 2.4 – Single guide RNAs used in tandem CRISPR/Cas9 constructs

Name	**5'-3' sequence**	**Final construct**
+204 sgRNA1	TTGGTTAGTCATGTGCGTCT	pX335 +204 4xsgRNA
+204 sgRNA2	CCAAAATGTCTGCAGCCAGT	pX335 +204 4xsgRNA
+204 sgRNA3	GCCTAGTTCACTTGAGCCTT	pX335 +204 4xsgRNA
+204 sgRNA4	CGCGTGTCCTTTCCTTGGAC	pX335 +204 4xsgRNA
+110 sgRNA1	CCCTTGAATGGGATTCATGC	pX335 +110 4xsgRNA
+110 sgRNA2	AGAGAGTCCTGATCTAAATG	pX335 +110 4xsgRNA
+110 sgRNA3	GTACTTTTCAGAGAGCAAGT	pX335 +110 4xsgRNA
+110 sgRNA4	AATTATAGTACGTGTCTGTC	pX335 +110 4xsgRNA
+23 sgRNA1	AGACCAACCAGTCCTAAGAT	pX335 +23 4xsgRNA
+23 sgRNA2	ACCAGAGGATCAGCACTTGG	pX335 +23 4xsgRNA
+23 sgRNA3	GTTGCTCAGTGGCCCCTTCT	pX335 +23 4xsgRNA
+23 sgRNA4	GGCATCCCCTTTATCGCCGA	pX335 +23 4xsgRNA
P2-CTCF sgRNA1	GACTGATCCTCGCGCCGTCG	pX459 P2-CTCF 2xsgRNA
P2-CTCF sgRNA2	AGCCCCGACATGACCGTGAA	pX459 P2-CTCF 2xsgRNA
P1-CTCF sgRNA1	GGGTCTCCATAGGGCAAGGC	pX459 P1-CTCF 2xsgRNA
P1-CTCF sgRNA2	GAGTCCTGTGATGATAGTCA	pX459 P1-CTCF 2xsgRNA
Prickle2 sgRNA1 D10A	GGTATGTAAAGCAAGGGACT	pX335 *Prickle2* 4xsgRNA
Prickle2 sgRNA2 D10A	ACTCCCTAGTAGCTAGCCGT	pX335 *Prickle2* 4xsgRNA
Prickle2 sgRNA3 D10A	GCCCACGTGTGAATATTTTA	pX335 *Prickle2* 4xsgRNA
Prickle2 sgRNA4 D10A	TGCATATTCCTGGAAGATAG	pX335 *Prickle2* 4xsgRNA
Prickle2 sgRNA1 Cas9	GAAATATTCACACGTGGGCA	pX459 *Prickle2* 2xsgRNA
Prickle2 sgRNA2 Cas9	GTACCAACGGCTAGCTACTA	pX459 *Prickle2* 2xsgRNA
Clic6-CTCF1 sgRNA1	ACAGCCGATTTTGTAGGCAA	pX459 *Clic6*-CTCF1 2xsgRNA
Clic6-CTCF1 sgRNA2	ACTTCAGATCAGCTTACCTT	pX459 *Clic6*-CTCF1 2xsgRNA
Clic6-CTCF3 sgRNA1	GTCCAACTGAAAACGCCCTT	pX459 *Clic6*-CTCF3 2xsgRNA
Clic6-CTCF3 sgRNA2	CTCTGTGACTAGGTCTCTGA	pX459 *Clic6*-CTCF3 2xsgRNA
Clic6-CTCF4 sgRNA1	CACACCGCCCCCCACTTCTT	pX459 *Clic6*-CTCF4 2xsgRNA
Clic6-CTCF4 sgRNA2	TACAGACCACCAGGAAAGTC	pX459 *Clic6*-CTCF4 2xsgRNA

Table 2.5 – Gibson assembly primers used to generate tandem CRISPR/Cas9 constructs

Name	5'-3' sequence	PCR template	Final construct
F1_tandem	GCCTTTTGCTGGCCTTTTGCTCACATG	pX_sgRNA_1	pX_4xsgRNA
R1_tandem	ATACCTATGCCGGATTGCCGAAAAAA	pX_sgRNA_1	pX_4xsgRNA
F2_tandem	CGCCAATCCGGCATAGGTATGAGGGC	pX_sgRNA_2	pX_4xsgRNA
R2_tandem	TCGCCACTCGATCCTGCCCGAAAAAA	pX_sgRNA_2	pX_4xsgRNA
F3_tandem	CGGGCAGGATCGAGTGGCGAGAGGA	pX_sgRNA_3	pX_4xsgRNA
R3_tandem	GCATATGTGCGTACCTAGCATAAAAAA	pX_sgRNA_3	pX_4xsgRNA
F4_tandem	ATGCTAGGTACGCACATATGCGAGGG	pX_sgRNA_4	pX_4xsgRNA
R4_tandem	GTAAGTTATGTAACGGGTACCAAAAAA	pX_sgRNA_4	pX_4xsgRNA
F1_tandem	GCCTTTTGCTGGCCTTTTGCTCACATG	pX_sgRNA_1	pX_2xsgRNA
R1_tandem	ATACCTATGCCGGATTGCCGAAAAAA	pX_sgRNA_1	pX_2xsgRNA
F2_tandem	CGCCAATCCGGCATAGGTATGAGGGC	pX_sgRNA_2	pX_2xsgRNA
R4_tandem	GTAAGTTATGTAACGGGTACCAAAAAA	pX_sgRNA_2	pX_2xsgRNA

2.7.2 CRISPR/Cas9 editing mESCs

Purified plasmids (Section 2.5.4) were transfected into mESCs using Lipofectamine 2000 (Invitrogen, 11668030) at a concentration of 5 μg per well of a 6-well plate. 24 hours before transfection, 1.5×10^6 cells were plated into a 6 well plate. 2 hours before transfection medium was refreshed. 9 μL Lipofectamine was added to 150 μL opti-MEM (Gibco, 31985062) in a tube, and 5 μg was added to 150 μL opti-MEM in a separate tube. Both tubes were combined, mixed, and incubated at RT for 5 mins. 290 μL of the transfection mixture was added drop-wise onto the cells in different locations. 24 hours after transfection, cells were either harvested and FACS sorted for GFP expression (pX335 transfected cells, see Section 2.2) or selected by adding puromycin to growth media at a final concentration of 1 μgmL^{-1}(Gibco, A1113803). Puromycin selection was performed for three days. 1×10^4 to 1×10^5 GFP-sorted cells were plated back directly in 24-well plates.

Three to seven days after selection and once cells had recovered, single colonies were picked according to an established protocol. Briefly, a limiting dilution of cells at 1:100, 1:200, 1:400, 1:800, 1:1600, 1:3200 were plated into duplicate 6-well plates. Seven to ten days after plating, when single colonies were visible to the naked eye, cells were washed in PBS and colonies manually picked using a P20 pipette and grown in flat bottom 96-well plates (Corning). Three days after colony picking, a crude DNA extraction was performed. Briefly, clones were harvested by trypsinisation for 5 mins followed by vigorous pipetting twenty times to generate a single cell suspension. 50% of the cells were replated in flat bottom 96-well culture plates for continued growth while the remaining cells were transferred to a 96-well PCR plate for DNA extraction. Cells for DNA extraction were centrifuged, supernatant removed, and pellet

resuspended in lysis buffer containing 50 mM Tris-HCl (pH 8.0, Sigma, T2694), 0.5% Triton X-100 (Sigma, 93443), 1 mgmL^{-1} Proteinase K (Thermo Fisher, EO0491). Lysis was performed at 65 °C for 1 hour, followed by enzyme inactivation at 95 °C for 10 mins. 2 μL crude DNA extracts were used in short-range (SR) PCR genotyping reactions (See section 2.7.4). Clones with a single knock-out (KO) band were then expanded up to 12 and 6 well plates before DNA was harvested using a DNeasy Blood and Tissue kit (Qiagen, 69504). Manufacturer's instructions were followed except DNA was eluted in 30 μL H_2O. After genotyping again using SR-PCR amplification from column extracted DNA, clones with a single KO band were cryopreserved (see Section 2.1.4) before further genotyping (see Section 2.7.4)

2.7.3 CRISPR/Cas9 editing 416B cells

1×10^7 416B cells were electroporated with 10 μg of pX335 plasmid using a Gene Pulser (Bio-Rad, 40mm cuvette, 220 mV, 960 μFD). GFP positive cells were FACS sorted (see Section 2.2) and centrifuged. Cells were resuspended in lysis buffer and crude DNA extraction was performed as in Section 2.7.2. 2 μL crude lysate were used in PCR as described in Section 2.7.4.

2.7.4 Mapping CRISPR/Cas9 deletions

Primers used in genotyping PCRs (Table 2.6) were designed using Primer Quest (IDT) and checked for specificity in the mouse genome (mm9) using UCSC BLAT (Kent, 2002) and UCSC in silico PCR. Optimum annealing temperatures for each primer pair were determined using a temperature gradient. SR and medium range (MR) PCRs were performed using HotStar Taq polymerase (Qiagen, 203445) Longer range (LR) PCRs were performed using Phusion polymerase (NEB, M0530S) as indicated in Table 2.6.

Table 2.6 – Primers used to genotype CRISPR/Cas9 knock-outs

Name	5'-3' sequence	T_a	Ext. Time	Polymerase
+204 SR F	GGGTTAGGAGTGTCCCTGATGT	63 °C	2 min	HotStar Taq
+204 SR R	GAGGCAGCTTTGGAAGGGAAAT	63 °C	2 min	HotStar Taq
+204 MR F	AGGTTGTCAGAACAAGCTACA	63 °C	3 min	HotStar Taq
+204 MR R	AGGGAGGAGTAAGGACGAATA	63 °C	3 min	HotStar Taq
+204 LR F	CATGACCAGGCCATTACAAATA	63 °C	5 min	Phusion GC
+204 LR R	ACCTGAGCCACATGAATAGA	63 °C	5 min	Phusion GC
+110 SR F	CCTCAGAGGCCAGAAGTACCTAAT	63 °C	2 min	HotStar Taq
+110 SR R	TTGGTGGTCTTTTCCCCGTG	63 °C	2 min	HotStar Taq
+110 MR F	GCTGCTCTCAACACAGTACAA	63 °C	3 min	HotStar Taq
+110 MR R	AGATCATTTGATGAAAGGCCTCA	63 °C	3 min	HotStar Taq
+110 LR F	CTCATGTGTACACGAGTGTTT	63 °C	5 min	Phusion GC

Continued

Table 2.6: continued from previous page

Name	5'-3' sequence	T_a	Ext. Time	Polymerase
+110 LR R	CAGGCATAAGAAGGGAGATTG	63 °C	5 min	Phusion GC
+23 SR F	TTCCTCCCATACCTCTCTCTGATATTG	63 °C	2 min	HotStar Taq
+23 SR R	GAGATGCTGGAGTTCCCTGAGATA	63 °C	2 min	HotStar Taq
+23 MR F	TGCTCAGCAGAATTGAGGTC	63 °C	3 min	HotStar Taq
+23 MR R	CTGCACAGCTCGGAGTATTT	63 °C	3 min	HotStar Taq
+23 LR F	GCTAGCAGGTGGAACCTAAA	63 °C	5 min	Phusion GC
+23 LR R	TCAAGGATCAGAGAAGCAAGAG	63 °C	5 min	Phusion GC
P2-CTCF SR F	CCTTACTTCTCTTGGGCCTTG	58 °C	1 min	HotStar Taq
P2-CTCF SR R	CTGGTGGCCACTTCCTAATG	58 °C	1 min	HotStar Taq
P2-CTCF MR F	GAAGTGGCACCGAGTCATTTA	63 °C	2 min	HotStar Taq
P2-CTCF MR R	CCTGATCGAGCTTCGAACTAAC	63 °C	2 min	HotStar Taq
P2-CTCF LR-3kb F	CTCGTTTGCATAGAGGAGAC	63 °C	3 min	Phusion GC
P2-CTCF LR-3kb R	CAGTTAGCCAGTCACGTAAG	63 °C	3 min	Phusion GC
P2-CTCF LR-5kb F	GCTACTAATGTATGTGCTCGT	63 °C	5 min	Phusion GC
P2-CTCF LR-5kb R	GCTCATGGTGTGTTAGAGTC	63 °C	5 min	Phusion GC
P1-CTCF SR F	ACTTAAGTGTCCACTCCGATTA	59 °C	1 min	HotStar Taq
P1-CTCF SR R	GGGATTAAGCACTTCTTTAGGC	59 °C	1 min	HotStar Taq
P1-CTCF MR F	CCACCTATTGACCTCTTCGTTC	63 °C	2 min	HotStar Taq
P1-CTCF MR R	TGCTACTGACTAATTTGAGGGTATT	63 °C	2 min	HotStar Taq
P1-CTCF LR-3kb F	CGAGCTCCACTCAAAGAAAT	63 °C	3 min	Phusion GC
P1-CTCF LR-3kb R	TCTAGGAAGGTCATGGAAATAAG	63 °C	3 min	Phusion GC
P1-CTCF LR-5kb F	TACTCACCTCTCATGAAGCA	63 °C	5 min	Phusion GC
P1-CTCF LR-5kb R	CCTTCTGCACAGAATGTCAA	63 °C	5 min	Phusion GC
Prickle2 SR F	CTCCGTAGCCTCTCCTAATGAT	63 °C	2 min	HotStar Taq
Prickle2 SR R	CTACACAGGAGTCTCTGAATCCAT	63 °C	2 min	HotStar Taq
Prickle2 MR F	TGCTGTAGCAAGCTTAGGGTGT	63 °C	3 min	HotStar Taq
Prickle2 MR R	GCGAAAAGACTGCCTCAAGTTC	63 °C	3 min	HotStar Taq
Clic6-CTCF1 SR F	TCCTTTCGTCCTCTGATGTA	63 °C	1 min	HotStar Taq
Clic6-CTCF1 SR R	CACCTGAGAGAACACAATACC	63 °C	1 min	HotStar Taq
Clic6-CTCF1 MR F	TCCATTGTCCACTGCTTTAC	63 °C	2 min	HotStar Taq
Clic6-CTCF1 MR R	GTATACAATTCTGCCCTAGCC	63 °C	2 min	HotStar Taq
Clic6-CTCF3 SR F	GTGGTCACTAGCTGTCTTCC	63 °C	1 min	HotStar Taq

Continued

Table 2.6: continued from previous page

Name	5'-3' sequence	T_a	Ext. Time	Polymerase
Clic6-CTCF3 SR R	GAAGTTGTGTAATGGCCACAG	63 °C	1 min	HotStar Taq
Clic6-CTCF3 MR F	GTATGCAGTTAGCCACAGAG	63 °C	2 min	HotStar Taq
Clic6-CTCF3 MR R	GCGGTGTGAATGATATCCAA	63 °C	2 min	HotStar Taq
Clic6-CTCF4 SR F	CCTTGTCAACAGCTCACTTT	63 °C	1 min	HotStar Taq
Clic6-CTCF4 SR R	CCTGCTCTAGCTTTGTTCTG	63 °C	1 min	HotStar Taq
Clic6-CTCF4 MR F	GCATATTTCCCTCTCAGAGTC	63 °C	2 min	HotStar Taq
Clic6-CTCF4 MR R	GAAATGGATAGGAGCAGGAAG	63 °C	2 min	HotStar Taq

PCR reactions using Phusion polymerase were performed by combining the following in a 0.2 mL tube and thermal cycling using the following protocol:

HotStar Taq PCR reaction components

10.0 μL HotStar 2x master mix (Qiagen, 203445)
6.0 μL Nuclease Free H_2O
1.5 μL FW primer [10 μM]
1.5 μL RV primer [10 μM]
1.0 μL 100 ngμL^{-1} gDNA]
Total: 20 μL

HotStar Taq PCR Thermal cycling conditions

a. 95 °C 10 mins
b. 95 °C 30 s
c. Ta (see Table 2.6) 30 s
d. 72 °C ext. T (see Table 2.6)
e. repeat steps b-d for a total of 40 cycles
d. 72 °C 10 min
g. 4 °C ∞

Phusion PCR reaction components

12.4 μL Nuclease Free H_2O
4 μL 5X Phusion GC buffer (NEB, M0530S)
0.4 μL dNTP mix [10 mM] (NEB, N0447S)
1.0 μL FW primer [10 μM]
1.0 μL RV primer [10 μM]
1.0 μL 100 ngμL^{-1} gDNA]
0.2 μL Phusion polyermase (NEB, M0530S)
Total: 20 μL

Phusion PCR Thermal cycling conditions

a. 98 °C 3 mins
b. 98 °C 10 s
c. Ta (see Table 2.6) 30 s
d. 72 °C ext. T (see Table 2.6)
e. repeat steps b-d for a total of 40 cycles
d. 72 °C 10 min
g. 4 °C ∞

To map deletions in isolated clones, SR PCR products were purified (Zymo, D4013) and Sanger sequenced (Source Bioscience). Deletions in pools of cells were isolated by TA cloning (Invitrogen, K202020) according to the manufacturer's instructions. Plasmids containing individual cloned deletion alleles were screened by restriction digestion using EcoRI (as in Section 2.5.1), purified as in Section 2.5.4, and Sanger sequenced.

2.7.5 Droplet digital PCR copy counting

Droplet Digital PCR-based quantification of larger deletions (LDs) in isolated clones using duplex EvaGreen reactions. A 200 bp amplicon (located on mouse chromosome 4 that was not targeted in any of these experiments) was used as an internal genomic control in every reaction. Several different 100 bp test amplicons were designed at periodic distances (100 bp, 500 bp, 1 kb, and 3 kb) away from CRISPR/Cas9 target sites. Each test and control amplicon combination was optimised to facilitate accurate distinction between two populations of positive droplets. Two or three 100 bp control amplicons were also included to allow normalisation to the diploid copy number. All primer sequences are located in Table 2.7. Reactions (22 μL) were performed as follows:

ddPCR reaction components

11 μL 2x QX200 ddPCR EvaGreen Supermix (Bio-Rad, 1864033)
8.35 μL Nuclease Free H_2O
0.55 μL 10 μM 200 bp Control Primer FW
0.55 μL 10 μM 200 bp Control Primer RV
0.275 μL 10 μM 100 bp Test Primer FW
0.275 μL 10 μM 100 bp Test Primer RV
1 μL gDNA [10-50 ng]
Total: 22 μL

Reaction mixes were set up in ddPCR 96-Well Plates (Bio-Rad, 12001925) and loaded into the Bio-Rad QX200 AutoDG and droplets were generated according to the manufacturer's instructions. Standard reagents and consumables were used, including cartridges and gaskets (Bio-Rad, 1864007), droplet generation oil (Bio-Rad, 1864005) and droplet reader oil (Bio-Rad, 18630004). Amplification was performed using a C1000 Touch™ Thermal Cycler.

Thermal cycling conditions

a. 95 °C 5 mins
b. 95 °C 30 s
c. 60 °C 1 min
d. repeat steps b-c for a total of 40 cycles
e. 4 °C 5 mins
f. 9 °C 5 mins
g. 4 °C ∞

Ramp rate set to 0.2 °Cs^{-1}

Post amplification, plates were loaded into the QX200 Droplet Reader (Bio-Rad). Relative concentrations of each test amplicon was (normalised to the quantity of the internal genomic control amplicon) was assessed using the QuantaSoft Analysis Pro™ software using at least 10,000 accepted droplets per sample. Ratios were calculated by applying Poisson statistics to the fraction of end-point positive reactions, and the 95% confidence interval of this measurement was used. Quantification of LDs in pools of cells was performed by our collaborators Alasdair Allan and Lydia Teboul at MRC Harwell institute. Reactions were performed as above except the target region flanking a CRISPR/Cas9 target site was amplified using a fluorescein amidite (FAM)-labelled assay selected from a Universal Probe Library (UPL) set (Human, sourced from Roche, Basel, SZ). Suitable probes and primers were identified using the ProbeFinder software at the Roche assay design centre (Table 2.7). In cases where a UPL set was not available, custom assays were ordered from LGC Biosearch Technologies (Novato, USA). UPL or custom assays were used in parallel with a VIC-labelled reference gene assay (Dot1l, Thermo Fisher) set at two copies (CNV2) on the Bio-Rad QX200 ddPCR System (Bio-Rad) as per Codner et al. (2018). Reaction mixes (22 μL) contained 100 ng of purified gDNA, 1× ddPCR Supermix for probes (Bio-Rad, 186-3026), 225 nM of each primer (two primers per assay) and 50 nM of each probe. Visualisation of the data was performed in R using ggplot2 taking the mean and 95% confidence intervals of two or three independent replicate transfections.

Table 2.7 – Droplet Digital PCR primers

Primer name	**Purpose**	**5'-3' sequence**
+204 5'-3kb F	Copy counting	AGACACTGACTGGATCTCTC
+204 5'-3kb R	Copy counting	AGAAGAACATTCTGACACAGG
+204 5'-1kb F	Copy counting	GGGAGGAGTAAGGACGAATA
+204 5'-1kb R	Copy counting	TAGAGTCTCGAAAGGCTGAG
+204 5-500bp F	Copy counting	GTGCCAACGGAGAATGAGA
+204 5-500bp R	Copy counting	GCTCTGTATTTAGGCAGAGTCC
+204 LOA F	Copy counting	CAAGGGTCCCAAATGAGAAG
+204 LOA R	Copy counting	TGAGTCAGGACACAGAGATG
+204 3'-500bp F	Copy counting	TCTATCGGTGACTGGAAGAG

Continued

Table 2.7: continued from previous page

Primer name	Purpose	5'-3' sequence
+204 3'-500bp R	Copy counting	TGAACAAAGGGTCTCACTAAC
+204 3'-1kb F	Copy counting	GGGTCAGACATTAAAGCACT
+204 3'-1kb R	Copy counting	GGGATTGGAGTTAACAGACC
+204 3'-3kb F	Copy counting	TGATGTTGGTACCTTTAAGACA
+204 3'-3kb R	Copy counting	AATGGAGGTCAAAGATAAGGG
+110 5'-3kb F	Copy counting	CCCAACAATGCTTTCTTTCC
+110 5'-3kb R	Copy counting	GTCTTACCAGGACACTTCAC
+110 5'-1kb F	Copy counting	CAACGTTAGCCACTCAGTAT
+110 5'-1kb R	Copy counting	AGTAGGGACTGTCAGTAAGG
+110 5-500bp F	Copy counting	GCCTCATCCTCTTCATCTG
+110 5-500bp R	Copy counting	CCCTCTCTCTCATCTCTCTT
+110 LOA F	Copy counting	GAGGCATTATCTGCCTTCAC
+110 LOA R	Copy counting	GATGGCAGTCTTGCATCAAC
+110 3'-1kb F	Copy counting	TCTCCCTGACCCTGAACTTT
+110 3'-1kb R	Copy counting	GTCAGGACTTCCAAGTGTAACC
+110 3'-3kb F	Copy counting	ATCTTCCAATGCTCTTGGTT
+110 3'-3kb R	Copy counting	CCCATGGTATAGAGAATGCAG
+23 5'-3kb F	Copy counting	GGATGCTTGTACCTACAGTTC
+23 5'-3kb R	Copy counting	ATGACCTAAGAAGCGAGAAAG
+23 5'-1kb F	Copy counting	CAAGAAAGGCATCAGACCAG
+23 5'-1kb R	Copy counting	TTTGTACTGTCAAGCCAAGG
+23 5-500bp F	Copy counting	GGTGGATGCTGGGAAGAG
+23 5-500bp R	Copy counting	GAAGCAGGGAAGGGATTTGA
+23 LOA F	Copy counting	CTTCTCAAAGAGCCTGGGAT
+23 LOA R	Copy counting	GCTTCAACTGCCGGTTTATT
+24 LOA F	Copy counting	GCTGGTATGACCACAAACTC
+24 LOA R	Copy counting	GCAGCTTTCCTGTTGATTCT
+23 3'-500bp F	Copy counting	CCCTTCATCACTGTCTGTGT
+23 3'-500bp R	Copy counting	CTCTAGCACTCTCCAAATCCC
+23 3'-1kb F	Copy counting	TCGTGAAAGAAGTCAAGTGG
+23 3'-1kb R	Copy counting	AATCAGTCTCTCAGCCTACA
+23 3'-3kb F	Copy counting	GGTAGGTACAGATGCTGTCT

Continued

Table 2.7: continued from previous page

Primer name	**Purpose**	**5'-3' sequence**
+23 3'-3kb R	Copy counting	GAGTCAGCATGATCCACAAG
P1-CTCF 5'-1kb F	Copy counting	GAGAGCAAAGGAGGAGAATTT
P1-CTCF 5'-1kb R	Copy counting	AGGCTACCAGTAGGCTTTAT
P1-CTCF 5-100bp F	Copy counting	TTCACTTCAGAAATTCCCAGAT
P1-CTCF 5-100bp R	Copy counting	GGCATCTAAACCATCCTTGT
P1-CTCF LOA F	Copy counting	GCTTTCAGGATAGCCTTCTG
P1-CTCF LOA R	Copy counting	ACTCCCGCCTTTGTAGT
P1-CTCF 3'-100bp F	Copy counting	GTTAAAGCCGCTCTCTAAGG
P1-CTCF 3'-100bp R	Copy counting	GGCTCTCAAGGAAGTGTAAA
P1-CTCF 3'-1kb F	Copy counting	GGCAGTTCTGACAGCTAAA
P1-CTCF 3'-1kb R	Copy counting	CAGAGAGGAAAGCTGGTAAC
P2-CTCF 5'-1kb F	Copy counting	CGAAACAGGAATCGAGAGAC
P2-CTCF 5'-1kb R	Copy counting	GGCAGCTTTGTGTCCAG
P2-CTCF 5-100bp F	Copy counting	CCCTCCACTTTCATCTGTG
P2-CTCF 5-100bp R	Copy counting	CCACAGCTTCTTCCTCTTC
P2-CTCF LOA F	Copy counting	AGGCCTTTCACGGTCAT
P2-CTCF LOA R	Copy counting	TTCCCAGGACGCCTAAC
P2-CTCF 3'-100bp F	Copy counting	GGGACAGACATTAGGAAGTG
P2-CTCF 3'-100bp R	Copy counting	CTACCACCGGTCTGAGAG
P2-CTCF 3'-1kb F	Copy counting	TCGAAGCTCGATCAGGATAA
P2-CTCF 3'-1kb R	Copy counting	TCAAGTGCGTGTCAAGTG
chrY *Sry* (Codner et al., 2016) F	Haploid control	CATCGGAGGGCTAAAGTGTC
chrY *Sry* (Codner et al., 2016) R	Haploid control	GTCCCACTGCAGAAGGTTGT
chrX Intergenic F	Haploid control	TCTGAAACTCCAGGAAGAGG
chrX Intergenic R	Haploid control	GTGGTCCAGATGGTGTTTAC
chr1 *Chpf* F	Diploid control	GGTAGTATCTCCACCGATGA
chr1 *Chpf* R	Diploid control	CTATACTGAAGCGCATGGAC
chr6 Intergenic F	Diploid control	TATAGGTGCAACCCGGAAT
chr6 Intergenic R	Diploid control	TCGGCCGACTAGTCTTTAG
chr7 *Prickle2* F	Diploid control	AGTAGCTAGCCGTTGGTAATAG
chr7 *Prickle2* R	Diploid control	GCACTGTGAGGGAAACAAAC
chr4 *Bach2* F	Internal control	GCAGGCTTGAGTAAGAAGAAG

Continued

Table 2.7: continued from previous page

Primer name	**Purpose**	**5'-3' sequence**
chr4 *Bach2* R	Internal control	GCAGGCTTGAGTAAGAAGAAG
+204 5'-3kb F	LD quantification	ACGCACGAGGAGAAGAACATG
+204 5'-3kb R	LD quantification	TCCACACTCTCAACTCACAGAAG
+204 5'-3kb Probe	LD quantification	CAGACACTGACTGGATCTCTCTTGCA
+204 5'-1kb F	LD quantification	TGACAGTGAATTTCTCCTCACC
+204 5'-1kb R	LD quantification	TCCAGCCTCATATTCGTCCT
+204 5'-1kb Probe	LD quantification	UPL52
+204 5-500bp F	LD quantification	CCACTCTGTCTGGAGGAGAGA
+204 5-500bp R	LD quantification	TGCAAGGGTTCAGGGATAAGTC
+204 5-500bp Probe	LD quantification	TGCCAACGGAGAATGAGAGGTGTTA
+204 5'-LOA F	LD quantification	GGATGCCTCCCTTCAAAGTAAC
+204 5'-LOA R	LD quantification	GCCAAGCATTTGTCTCCCA
+204 5'-LOA Probe	LD quantification	CGTTGGAACAGTGACCGCAGCC
+204 LOA F	LD quantification	GCCACCACTTTCCTACACAA
+204 LOA R	LD quantification	GGAGGCTTGCACCTTCTCAT
+204 LOA Probe	LD quantification	ATCCTGGGTCTTTCTTCACAAGGGTC
+204 3'-LOA F	LD quantification	CCACCAAAGGCTCAAGTGAAC
+204 3'-LOA R	LD quantification	CCCTCCCTGTCCAAGGAAA
+204 3'-LOA Probe	LD quantification	TAGGCTGATGTCCCCGCGTGT
+204 3'-500bp F	LD quantification	CTGTGGCCTCAAGTGTAGTC
+204 3'-500bp R	LD quantification	GGCCGCTCCTTTCTTCTGA
+204 3'-500bp Probe	LD quantification	TGCCTTAGCGAGTGATCTGTGTGC
+204 3'-1kb F	LD quantification	CCTGTTGCACCTGGTTCTTAG
+204 3'-1kb R	LD quantification	TGAAGTCACATAAAATGTTTCATTCC
+204 3'-1kb Probe	LD quantification	UPL16
+204 3'-3kb F	LD quantification	GCAGGCCACTCTGGTATGAA
+204 3'-3kb R	LD quantification	GGAGACAGTACAAAGCACCATGAC
+204 3'-3kb Probe	LD quantification	TGGCAGCGTCATGGTGGACTCT
Prickle2 5'-3kb F	LD quantification	AGCACACAACAGTCCTCTCA
Prickle2 5'-3kb R	LD quantification	CCACTACAAGCATGGGACAG
Prickle2 5'-3kb Probe	LD quantification	CGAAAGCTCTTCCTGTCATTGTCTTTATGA
Prickle2 5'-1kb F	LD quantification	CCTAGGCGGCTAACCTCAC

Continued

Table 2.7: continued from previous page

Primer name	Purpose	5'-3' sequence
Prickle2 5'-1kb R	LD quantification	GAGAAGGGTAGGCAATGGTG
Prickle2 5'-1kb Probe	LD quantification	UPL52
Prickle2 5-500bp F	LD quantification	GACCTCATGACCAATACCATGA
Prickle2 5-500bp R	LD quantification	GTTCTGAGGGTCTTGGAGGA
Prickle2 5-500bp Probe	LD quantification	UPL63
Prickle2 5'-LOA F	LD quantification	GCCGTTGGTAATAGTTGGATCATTC
Prickle2 5'-LOA R	LD quantification	CGCACTGTGAGGGAAACAAAC
Prickle2 5'-LOA Probe	LD quantification	TCACTATTGTGAGAGCTCCCGCTG
Prickle2 LOA F	LD quantification	AGTGCGGAGGGCAGATTAAGG
Prickle2 LOA R	LD quantification	AGGGTGGATGCCAGCAGATG
Prickle2 LOA Probe	LD quantification	CGGCGACATCGCTGTGTTTGCAT
Prickle2 3'-LOA F	LD quantification	TGCAGCCTGTGATGAGGTTAG
Prickle2 3'-LOA R	LD quantification	CTGAAGGTGTCACCCCTATCTTC
Prickle2 3'-LOA Probe	LD quantification	AGCCCCTGCCCACGTGTGAATAT
Prickle2 3'-500bp F	LD quantification	GCTCATGACCTTTAACTTCCTACAC
Prickle2 3'-500bp R	LD quantification	TTTCTGAAAAGGGTAACATAAGTGAGT
Prickle2 3'-500bp Probe	LD quantification	UPL16
Prickle2 3'-1kb F	LD quantification	CGGGCTGAGCTTATAAATCG
Prickle2 3'-1kb R	LD quantification	CCTGGTGGATCTGACAGACA
Prickle2 3'-1kb Probe	LD quantification	UPL38
Prickle2 3'-3kb F	LD quantification	TGCATAGATGCGATGTTGTTT
Prickle2 3'-3kb R	LD quantification	CCACCGTGGAGGAAGTTACA
Prickle2 3'-3kb Probe	LD quantification	UPL34

2.7.6 Targeted next-generation sequencing

Locus specific next-generation sequencing primers (Table 2.8) were designed to amplify a 225 bp amplicon centred on each CRISPR/Cas9 target site. Primers were modified to contain Illumina® Truseq adapter sequences at the 5' end. PCR was performed (25 cycles, Herculase II PCR kit (Agilent)) on a pool of gDNA harvested from cells 3-7 days post transfection. Truseq indices (NEB, E7335) were added to the PCR amplified fragments by using a further 6 cycles of PCR with Herculase II PCR kit (Agilent, 600675) according to the manufacturer's instructions with annealing temperature set to 63 °C and extension time of 1 min.

Table 2.8 – Primers used for deep sequencing of short deletions

Name	5'-3' sequence
Prickle2 FW	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAGGCAGAGATGACACTGAA
Prickle2 RV	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCAATGAGCTTCTGGTTGAC
Clic6-CTCF3 FW	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCACTGAAATCTGGTGCAT
Clic6-CTCF3 RV	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTTGCCCTCAGAGACCTAGT

2.8 Chromatin assays

2.8.1 RNA-seq

Duplicate samples of 2.5×10^6 E14, 416B, and HPC7 cells were collected, washed once in PBS, and resuspended in 700 μL QIAzol Lysis Reagent (Qiagen, 79306). Cells in QIAzol were vortexted thoroughly for one minute and placed at room temperature for 5 minutes. Cell pellets were snap frozen on dry ice and stored at -80ºC for less than one month. Total RNA was extracted using the miRNeasy Mini kit (Qiagen, 217004) according to the manufacturer's instructions. DNaseI digestion was performed on-column according to the manufacturer's instructions. RNA quantity was determined using a Qubit™ RNA BR Assay kit (Invitrogen, Q10210). RNA integrity was determined using a 4200 TapeStation System (Agilent) and RNA ScreenTape and reagents (Agilent, 5067-5576, 5067-5577, 5067-5578) according to the manufacturer's instructions. All RNA samples had a RINe score of 10.0. Ribosomal RNA was depleted from 2.5 μg total RNA per sample using the RiboMinus™ Eukaryote System v2 (Invitrogen, A15026) according to the manufacturer's instructions. Depletion was confirmed using RNA ScreenTape again. Poly A plus RNA was selected using the NEBNext® Ultra directional library prep kit (NEB, E7420) and poly A selection module (NEB, E7490) according to the manufacturer's instructions. Poly A minus RNA was isolated from poly A plus mRNA by at each step keeping all of the supernatant after separating poly-A plus mRNA bound to the beads from solution that the protocol recommends discarding. Poly A minus RNA was purified from the combined supernatants using Ampure RNAClean XP (Beckman Coulter, A63987). Poly A minus RNA was subject to a final round of poly A plus mRNA depletion using the poly A selection module (NEB, E7490), purified again using Ampure RNAClean XP before eluting in 15.5 μL First Strand Synthesis Reaction Buffer and Random Primer mix (2 x) from the NEBNext® Ultra directional library prep kit. Synthesis of cDNA from poly A plus and minus RNA was performed using the NEBNext® Ultra directional library prep kit according to the manufacturer's instructions. Adapter ligation and 15 cycles of PCR enrichment was performed exactly according to the NEBNext® Ultra directional library prep kit protocol. Quality control of sequencing libraries was performed as in Section 2.8.7 and libraries were sequenced on a 75 cycle High output Illumina® NextSeq run.

2.8.2 ATAC-seq

2.8.2.1 Transposition reaction

ATAC-seq libraries were generated from sorted populations of 2-5×10^4 differentiated mESCs as previously described (Buenrostro et al., 2013) with minor modifications to facilitate library preparation from low cell numbers. Sorted cells were centrifuged at 500 rcf for 15 min at 4 °C. Supernatant was removed and cells were resuspended in cold lysis buffer (10 mM Tris-HCl, pH 7.4 [Sigma, T2194], 10 mM NaCl [Sigma, S9625], 3 mM $MgCl_2$ (Sigma, M1028) and 0.1% IGEPAL CA-630 [Sigma, I8896]) and incubated on ice for 10 min. Cells were centrifuged at 500 rcf for 15 min at 4 °C and supernatant removed and discarded. Nuclei were carefully resuspended in 10 μL of transposition reaction mix (containing 5 μL 2x TD Buffer (Illumina, FC-121-1030), 0.35 μL Tn5 transposase (Illumina, FC-121-1030), and 4.65 nuclease free H_2O) by pipetting gently. Transposase reactions were incubated for 30 mins at 37 °C and then immediately quenched by the addition of 1.1 μL 500 mM EDTA. Reactions were mixed gently by flicking and then centrifuged for 1 min at 500 rcf at RT. Reactions were then incubated at 50 °C for 10 mins and carried straight through to PCR amplification.

2.8.2.2 PCR amplification, purification, and quality control

An additional 10 μL 50 mM $MgCl_2$ was added to each reaction to ensure sufficient Mg^{2+} ions was available for PCR amplification. PCR reactions were set up as follows using custom nextera primer sequences that are available in Table 2.9.

ATAC-seq library PCR amplification

21.1 μL Transposed DNA
2.65 μL Nuclease Free H_2O
0.625 μL 25 μM Customized Nextera PCR Primer 1 (Ad1_noMX)
0.625 μL 25 μM Customized Nextera PCR Primer 2 (Ad2_barcode)
Total: 25 μL

Thermal cycling conditions

a. 72 °C 5 mins
b. 98 °C 30 s
c. 98 °C 10 s
d. 63 °C 30 s
e. 72 °C 1 min
f. repeat steps c-e for a total of 13 cycles
g. 4 °C ∞

Table 2.9 – Primers used in ATAC-Seq

Name	5'-3' sequence
Ad1_NoMx	AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG
Ad2.1	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGT
Ad2.2	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGT
Ad2.3	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGT
Ad2.4	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGT
Ad2.5	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGT
Ad2.6	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGT
Ad2.7	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGATGT
Ad2.8	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGGAGATGT
Ad2.9	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGGAGATGT
Ad2.10	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGGAGATGT
Ad2.11	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGGAGATGT
Ad2.12	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGGAGATGT
Ad2.13	CAAGCAGAAGACGGCATACGAGATATCACGACGTCTCGTGGGCTCGGAGATGT
Ad2.14	CAAGCAGAAGACGGCATACGAGATACAGTGGTGTCTCGTGGGCTCGGAGATGT
Ad2.15	CAAGCAGAAGACGGCATACGAGATCAGATCCAGTCTCGTGGGCTCGGAGATGT
Ad2.16	CAAGCAGAAGACGGCATACGAGATACAAACGGGTCTCGTGGGCTCGGAGATGT
Ad2.17	CAAGCAGAAGACGGCATACGAGATACCCAGCAGTCTCGTGGGCTCGGAGATGT
Ad2.18	CAAGCAGAAGACGGCATACGAGATAACCCCTCGTCTCGTGGGCTCGGAGATGT
Ad2.19	CAAGCAGAAGACGGCATACGAGATCCCAACCTGTCTCGTGGGCTCGGAGATGT
Ad2.20	CAAGCAGAAGACGGCATACGAGATCACCACACGTCTCGTGGGCTCGGAGATGT
Ad2.21	CAAGCAGAAGACGGCATACGAGATGAAACCCAGTCTCGTGGGCTCGGAGATGT
Ad2.22	CAAGCAGAAGACGGCATACGAGATTGTGACCAGTCTCGTGGGCTCGGAGATGT
Ad2.23	CAAGCAGAAGACGGCATACGAGATAGGGTCAAGTCTCGTGGGCTCGGAGATGT
Ad2.24	CAAGCAGAAGACGGCATACGAGATAGGAGTGGGTCTCGTGGGCTCGGAGATGT

PCR reactions were purified using MinElute PCR purification kit (Qiagen, 28004) and eluted in 15 μL elution buffer.

2.8.3 Next Generation Capture-C

NG Capture-C was performed as previously described (Davies et al., 2015) and as detailed below.

2.8.3.1 3C library generation

3C libraries were generated from 1×10^7 $1x10^7$ 416B or E14 cells and three replicates for each cell type were generated on different days. Cultured cells were crosslinked using 2% formaldehyde (Sigma, 252549) for 10 mins at RT and quenched by adding 1.5 mL Glycine (G1726) 1 M solution in H_2O (Sigma, W4502) at 4 °C. Cells were lysed and cell pellets frozen at -80 °C. Cell pellets were thawed, and homogenised in DpnII restriction enzyme buffer (NEB, R0543M) using a Dounce homogeniser. DpnII digestion was performed in an Eppendorf ThermoMixer®shaking at 1400rpm overnight at 37 °C. Ligation was performed on heat-inactivated digests using HC T4 DNA ligase (Life Technologies, EL0013) shaking at 1400rpm overnight at 16 °C. Digestion reactions were decrosslinked using Proteinase K (Thermo Fisher, EO0491) overnight at 65 °C and DNA was extracted by phenol-chloroform-isoamylalcohol 25:24:1 (Sigma, 77617). Digestion efficiency was calculated by qPCR and only libraries with greater than 80% digestion efficiency were used for next generation sequencing library generation.

2.8.3.2 Next-generation sequencing library generation

NEBNext® DNA Library Prep Reagent Set for Illumina® (NEB, E6000) was performed according to the manufacturer's instructions. 3C libraries were sonicated using a Covaris ultrasonicator (duty cycle 10%, intensity 5, cycles per burst 200, time 360s, set mode frequency sweeping). 1.8x volume Agencourt Ampure XP SPRI beads (Beckman Coulter, A63881) were used to purify DNA according to the manufacturer's instructions after each of the following library preparation steps. Sonicated libraries were end repaired for 30 mins at 20 °C in a thermal cycler. dA-tailing was performed for 30 mins at 37 °C in a thermal cycler. Adaptor ligation was performed for 15 mins at 20 °C in a thermal cycler followed by addition of USER™ enzyme and incubation for 30 mins at 37 °C in a thermal cycler. PCR addition of indices was performed using Herculase II Fusion polymerase kit (Agilent, 600677) according to the manufacturer's instructions for 7 cycles.

2.8.3.3 Capturing interacting sequences

Capture probes were designed against DpnII fragments containing desired regions of interest. 120bp 5' biotinylated oligonucleotides (IDT) were used that had a BLAT density score of less than 23 when analysed using CapSequm. Probes were also checked using UCSC BLAT to make sure that they had an insignificant number of minor scoring matches when analysed against the mouse genome (mm9). Probe names and coordinates are located in Table 2.10.

Table 2.10 – Capture-C Oligonucleotides

Oligo name	Coordinates mm9
Rcan1_a	chr16 + 92466149 92466268
Rcan1_b	chr16 + 92467229 92467348
Clic6_a	chr16 + 92497541 92497660
Clic6_b	chr16 + 92498906 92499025
CTCF_+380_a	chr16 + 92445723 92445842
CTCF_+314_a	chr16 + 92511622 92511741

Continued

Table 2.10: continued from previous page

Oligo name	Coordinates mm9
CTCF_+314_b	chr16 + 92512874 92512993
CTCF_+260_a	chr16 + 92565147 92565266
CTCF_+260_b	chr16 + 92565638 92565757
+205_a	chr16 + 92620938 92621057
+172_a	chr16 + 92653803 92653922
+172_b	chr16 + 92654565 92654684
+110_a	chr16 + 92715968 92716087
+23_a	chr16 + 92801788 92801907
+23_b	chr16 + 92802295 92802414
-59_a	chr16 + 92884103 92884222
-59_b	chr16 + 92884989 92885108
CTCF_-468_a	chr16 + 93293638 93293757
CTCF_-468_b	chr16 + 93294897 93295016
CTCF_-629_a	chr16 + 93454798 93454917
CTCF_-718_a	chr16 + 93542965 93543084
CTCF_-718_b	chr16 + 93544698 93544817
Setd4_a	chr16 + 93604349 93604468
Cbr1_a	chr16 + 93607029 93607148
Cbr1_b	chr16 + 93608154 93608273
Cbr3_a	chr16 + 93683513 93683632
Dopey2_a	chr16 + 93712643 93712762
Runx1_P1	chr16 + 92825340 92825459
Runx1_P2a	chr16 + 92697424 92697543
Runx1_P2b	chr16 + 92697750 92697869

Pooled capture oligonucleotides were used at a concentration of 2.9nM each. A pooled hybridisation mixture containing 2 μg indexed 3C library, 5 μg mouse COT-1 DNA (Invitrogen, 18440016), 1 nmol TS-HE Universal Oligo (Roche, 06777287001), 1 nmol (each) TS-HE Index Oligo (Roche, 06777287001), was dried using a vacuum centrifuge before addition of 7.5μL hybridisation buffer and 3μL Component A (Roche, 07145594001). The hybridisation mixture was denatured for at 95°C for 10 mins before addition to capture probes preheated to 47°C in a thermal cycler. The capture reaction was performed 72hours. Washing and recovery of captured material using Nimblegen SeqCap EZ hybridisation and wash kit (Roche, 05634261001) according to the manufacturer's instructions and captured sequences

were pulled down using M-270 Streptavidin Dynabeads (Invitrogen, 65305). PCR amplification was performed by amplifying directly from the Dynabeads using KAPA HiFi HotStart ReadyMix (Roche, 07958897001) and POST-LM_PCR Oligo 1&2 (Roche, 07145594001) for 12 cycles according to the manufacturers instructions. Amplfied DNA was purified using 1.8x volume Ampure XP SPRI beads. A second Capture and PCR cycle was performed in order to further enrich for captured interactions.

2.8.4 Tiled Capture-C

Tiled Capture-C oligos were designed by Marieke Oudelaar and synthesised in-house by Sara De Ornellas. Oligos were modified with a 5' biotin and were 70 nt in length. A total of 7440 oligos could be designed with a BLAT density score of less than 30. The region covered was 2.7 Mb over mm9 coordinates chr16 91567700 94100269. Oligos were not diluted before setting up pooled hybridisation reactions using 1 μg of indexed 3C libraries. All other steps were performed identically to Next-Generation Capture-C as detailed in Section 2.8.3.

2.8.5 ChIP-seq

2.8.5.1 Cell fixation

For CTCF ChIP cells were counted and 1×10^6 cells were resuspended in 900 μL culture medium. 100 μL of fixative solution (containing 50 mM HEPES pH 8.0 [Gibco, 15630-056], 1 mM EDTA pH 8.0 [Gibco, 15575-038] , 0.5 mM EGTA pH 8.0 [Sigma, E3889], 100 mM NaCl [Sigma, S9625] and 10% formaldehyde [Sigma, 252549]) was added to the medium to reach a final formaldehyde concentration of 1%. Cells were incubated for 10 mins at RT on a rotating platform before quenching with 150 μL 1 M cold glycine (Sigma, G1726) and a further 5 mins incubation rotating at RT.

For Rad21 ChIP cells were crosslinked using 2 mM disuccinimidyl glutarate (DSG, Sigma, 80424) and 1% paraformaldehyde. Briefly, cells were counted and 1×10^7 cells were resuspended in 1 mL 2 mM DSG in PBS and cells were incubated for 30 mins at RT on a rotating platform. Cells were centrifuged and supernatant discarded. Cells were resuspended in 1 mL 1% formaldehyde in PBS and incubated for 30 mins at RT on a rotating platform.

Cells were centrifuged, washed in 500 μL cold PBS, centrifuged again, supernatant was discarded and cells were snap frozen in a dry ice and ethanol (VWR, 20821.330) bath. Frozen cell pellets were stored for up to several months at -80 ^{o}C.

2.8.5.2 Lysis and sonication

CTCF ChIP steps were conducted using Millipore ChIP agarose kit (Millipore, 17-295) mostly following its protocol. Millipore ChIP magnetic bead kit (Millipore, 17-10085) was used for Rad21 ChIP mostly following its protocol.

For CTCF ChIP, cell pellets were thawed on ice before being resuspended using 120 μL lysis buffer (Millipore, 17-295) with the addition of cOmplete Protease Inhibitor Cocktail (PIC, Sigma, 11873580001). PIC was prepared by dissolving 1 tablet in 2 mL nuclease free H_2O to make a 25X concentrated solution. Cells in lysis buffer were incubated on ice for 15 mins with mixing by pipetting every 5 mins. For Rad21 ChIP, thawed cell pellets were resuspended in 500 μL cell lysis buffer (Millipore, 17-10085) with the addition of PIC. Cells were incubated for 15 mins on ice and mixed by pipetting every 5 mins. Cells were centrifuged for 5 mins at 600 rcf at 4 ^{o}C and then resuspended in 120 μL nuclear lysis buffer with the addition of PIC and incubated on ice for a further 5 mins.

Lysed cells were transferred to a Covaris microTUBE AFA Fiber pre-split snap-cap 6x16mm (Covaris, 520045) being careful not to introduce any bubbles. For single crosslinked (CTCF ChIP) samples the following protocol was used: Duty Cycle: 2% Intensity: 3.0, Cycles/burst: 200, Power mode: Freq. Sweeping, Duration: 120s, Temp: 6°C, Batches: 4. For double crosslinked (Rad21 ChIP) samples the following protocol was used: Duty Cycle: 10% Intensity: 5.0, Cycles/burst: 200, Power mode: Freq. Sweeping, Duration: 300s, Temp: 6°C Batches: 2.

2.8.5.3 Immunoprecipitation and recovery

Sonicated chromatin was transferred to a microcentrifuge tube and covaris microtube was washed with 120 μL dilution buffer (Millipore, 17-295, 17-10085) with the addition of PIC and combined with the sonicated chromatin. Sonicated chromatin was centrifuged for 10 mins at 20000 rcf at 4 °C and supernatant was transferred to a new tube. For Rad21 ChIP, sonicated chromatin was pre-cleared by adding 10 μL thoroughly resuspended protein A/G magnetic bead slurry (Millipore, 17-10085) and incubated for 5 mins on a rotating platform at 4 °C. Chromatin was centrifuged for 1 min at 1000 rcf at 4 °C to pellet beads and supernatant was removed. Sonicated chromatin was diluted to 1.2 mL using dilution buffer with the addition of PIC and 50 μL was removed as the 5% input control, for Rad21 and CTCF ChIP, respectively. Antibodies against CTCF (EMD Millipore, 07-729) and Rad21 (Abcam, ab992) were added to 1 mL chromatin (5 μLmL^{-1} anti-CTCF and 2 μLmL^{-1} anti-Rad21) and incubated on a rotating platform overnight at 4°C. The next morning, 20 μL protein A/G magnetic bead slurry (Rad21 ChIP) or 60 μL salmon sperm / Protein A agarose slurry (CTCF ChIP) was added to the chromatin and incubated for 1 (CTCF ChIP) or 5 (Rad21 ChIP) hours on a rotating platform at 4 °C.

Beads were seperated from unbound chromatin using either a Dyna-Mag-2 magnet (Thermo Fisher, 12321D) for Rad21 ChIP, or centrifugation for 1 min at 1000 rcf at 4 °C for CTCF ChIP. Washes were performed by resuspending the beads in 500 μL of wash buffer with the addition of PIC and then incubated for 5 mins on a rotating platform at 4 °C before pelleting using either magnetic isolation or centrifugation (see above). Successive washes were performed in the following order: Low Salt Wash Buffer, High Salt Wash Buffer, LiCl Wash Buffer, TE Buffer. After the first wash step beads were transferred to a fresh tube to reduce background.

For CTCF ChIP, bound fragments were eluted from agarose beads by resuspension in 250 μL fresh elution buffer (containing 1% SDS, 0.1 M $NaHCO^3$ [Sigma, S8761]) and incubating at RT for 15 min. Beads were centrifuged for 1 min at 1000 rcf at RT and eluate was transferred to a fresh tube. The elution step was repeated and 500 μL eluate combined in a fresh tube with the addition of 20 μL 5 M NaCl (Sigma, S9625). Input samples were removed from storage at -20 °C and diluted with elution buffer to 500 μL and 20 μL 5 M NaCl was added. Immunoprecipitates and input samples were incubated at 65 °C over night. The next morning, 10 μL EDTA, 20 μL 1 M Tris-HCl pH 6.5 [Millipore, 17-295] and 2 μL proteinase K (10 mgmL^{-1}, Thermo Fisher, EO0491) were added and samples incubated at 45 °C for 1 hour.

For Rad21 ChIP, beads were centrifuged for 3 mins at 960 rcf at RT and supernatant removed. Beads were resuspended in 105 μL of elution buffer (consisting of 50 mM Tris-HCl,pH 8.0, 10 mM EDTA, 1% SDS) and incubated at 65 °C for 30 mins. Beads were centrifuged for 1 min at 16000 rcf at RT and 100 μL supernatant was transferred to a new tube. Input was diluted to 100 μL using elution buffer and both were incubated at 65 °C over night. The next morning, 1 μL RNAse A (500 μgmL^{-1}, Thermo Fisher, 12091021) was

added and samples incubated at 37 °C for 30 mins. 2 μL proteinase K was added, and incubated at 45 °C for 30 mins. Samples were incubated at 95 °C for 10 mins and then cooled to RT.

2.8.5.4 DNA purification

DNA purification was performed by adding 1x volume phenol-chloroform-isoamylalcohol (25:24:1, Sigma, 77617) and vortexing each sample thoroughly for 10 seconds. DNA mixture was then transferred to a light PhaseLock gel tube (5Prime, 733-2477) and centrifuged for 10 mins at 15000 rcf at RT. Upper layer was transferred to a new tube and DNA was precipitated by adding 0.1x volume μL 3 M NaOAc pH 5.5 (Invitrogen, AM9740), 2 μL GlycoBlue (Invitrogen, AM9515), 3x volume 100% ethanol, and incubating at -80 °C for a mininum of 2 hours. DNA was pelleted by centrifugation for 90 mins at 21000 rcf at 4 °C. Supernatant was removed, pellet was washed with 500 μL ice cold 70% ethanol, and centrifuged again for 5 mins at 21000 rcf at 4 °C. Supernatant was removed completely and pellet was allowed to air dry for 5 mins. DNA was resuspended in 50 μL H_2O using an Eppendorf ThermoMixer® set to 56 °C 800 rpm, for 30 mins.

2.8.5.5 Quality control of sonication and enrichment

Sonication quality and fragment size was determined by analysing 1 μL of input samples on a D1000 TapeStation (Agilent, 5067- 5582). Input and immunoprecipitated DNA samples were quantified using Qubit dsDNA BR Assay Kit (Thermo Fisher, Q32850) and Qubit dsDNA HS Assay Kit (Thermo Fisher, Q32851), respectively. Enrichment was determined using qPCR at two positive control sites known to be bound in similar cell types and one negative control region. Sequenced ChIP libraries were enriched 13 to 130 fold. Primer sequences are located in Table 2.11. 15 μL qPCR reactions were set up in technical duplicates using KAPA Fast KAPA SYBR FAST qPCR Master Mix (Sigma, KK460) according to the manufacturer's instructions.

Table 2.11 – Primers used for ChIP enrichment

Name	5'-3' sequence
CTCF Chpf FW	GGCCAAGATAGAGATGGGTTG
CTCF Chpf RV	GTGCCTGATGCCACCTATAC
CTCF Cpn2 FW	CAATTACCAACTCCGTTCCT
CTCF Cpn2 RV	ATTGGTAGAACGACTTTCCG
CTCF chr7 FW	TATAGGTGCAACCCGGAAT
CTCF chr7 RV	TCGGCCGACTAGTCTTTAG
mouse neg chr16 a FW	CAGGAACACACAGTGAAGAA
mouse neg chr16 a RV	CCTGTTTGGCCAAGTCTAA
mouse neg chr16 b FW	AAGGCTGAAATGCGGATAAA
mouse neg chr16 b RV	CCACTTTCCAGCTCTAGGTA
mouse neg chr1 a FW	GGAATGATCACGTCCTTAGC
mouse neg chr1 a RV	GAGATGTGTTGGGTCTGTAAA

Enrichment and percentage input was calculated using the following equations:

$$\Delta C_{\mathrm{t}} = C_{\mathrm{t}}(5\%\mathrm{Input}) - C_{\mathrm{t}}(\mathrm{ChIP})$$

$$\%\ \mathrm{input} = \frac{2^{\Delta C_{\mathrm{t}}}}{20 \times 100}$$

$$\mathrm{Enrichment} = \frac{\%\ \mathrm{input\ positive\ site}}{\%\ \mathrm{input\ negative\ site}}$$

2.8.5.6 Indexing

NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (NEB, E7645S) was used according to the manufacturer's instructions. 0.9x volume Agencourt Ampure XP SPRI beads (Beckman Coulter, A63881) were used to purify DNA according to the manufacturer's instructions. DNA was eluted in 15 μL 0.1x TE buffer (Thermo Fisher, AM9849) and PCR amplification was performed using NEBNext® Q5 HotStart (NEB, E7645S) for 11 cycles according to the manufacturer's instructions. Post amplification, 0.9x volume Agencourt Ampure XP SPRI beads were used to purify DNA which was eluted in 30 μL H_2O.

2.8.6 Bisulfite-seq

Genomic DNA (gDNA) was extracted from 1-5$\times 10^6$ cells using a Qiagen DNeasy Blood and Tissue Kit (Qiagen, 69504) according to the manufacturer's instructions. 250 ng gDNA, or Universal Methylated Mouse DNA Standard (Zymo, D5012) was bisulfite converted by Akin Bucakci using EZ DNA Methylation-Gold™ Kit (Zymo, D5005) according to the manufacturer's instructions and eluted in 10μL volume. Several nested PCR primer sets were designed by Danuta Jeziorska to amplify 281-379 bp overlapping target regions of the mm9 genome (primers listed in Table 2.12). 10 μL first methylation specific PCRs were performed by Akin Bucakci on 1 μL bisulfite converted DNA using HotStarTaq DNA Polymerase (Qiagen, 203205) and Fo Ro (outside) primer pairs. Second nested PCR reactions were performed using the same reaction conditions but using 1 μL of the first PCR reaction as a template (instead of bisulfite converted DNA) and using Fi Ri (inside) primer pairs.

PCR amplification

5.5 μL H_2O
1.0 μL Buffer
0.2 μL dNTPs
0.4 μL Primers [25 μM each]
0.1 μL HotStart Polymerase
1.0 μL Bisulfite converted DNA
Total: 10 μL

Thermal cycling conditions

a. 95 °C 15 mins
b. 95 °C 1 min
c. 52-55 °C 2 mins

d. 72 °C 3 mins
e. repeat steps b-d for a total of 13 cycles
f. 72 °C 10 mins
g. 4 °C ∞

Table 2.12 – Primers used for Bisulfite sequencing

Name	5'-3' sequence
djRunx1 1Fo	ATGGGTAGGGTTTTGTTGTAG
djRunx1 1Fi	GTAGAGGAAGTTGGGGTTG
djRunx1 1Ri	CTAAACTTATATATTAATCTC
djRunx1 1Ro	CTAATAATAATTTAATCTCCTACC
djRunx1 2Fo	GATTTTTTGTAAGTTGTTTT
djRunx1 2Fi	GATTAATATATAAGTTTAGAAG
djRunx1 2Ri	CATTTTTTAATTATTATTATTATTC
djRunx1 2Ro	CAAACTAACAAACAACTAAC
djRunx1 3Fo	GAGAGTTGAAGTTTAGTT
djRunx1 3Fi	GAATAATAATAATAATTAAAAAATG
djRunx1 3Ri	CAACAAACATTAATAAATTTC
djRunx1 3Ro	ACTAATTTTCAAAACTAC
djRunx1 4Fo	GTAGAGAATTTAAATATAATAGaa
djRunx1 4Fi	GAAATTTATTAATGTTTGTTG
djRunx1 4Ri	CAATATAAACAATAACTAAC
djRunx1 4Ro	CATCTTACCAAACTCTAAAC
djRunx1 5Fo	GAAGTTAGTTATTGTTTATATTG
djRunx1 5Fi	GTTTAGAGTTTGGTAAGATGA
djRunx1 5Ri	ccCTTTCTTACTTTTAAATAAAC
djRunx1 5Ro	AACCTTCTTATTTTCCTTTC
djRunx1 6Fo	GTTTATTTAAAAGTAAGAAAG
djRunx1 6Fi	GAAAGGAAAATAAGAAGGTTT
djRunx1 6Ri	AAAAAACCTATAATTAAC
djRunx1 6Ro	TCTAAACCTAAACCTAAC
djRunx1 7Fo	AGGAATTTTATTTTTTTAGTT
djRunx1 7Fi	GTTAATTATAGGTTTTTT
djRunx1 7Ri	CTAATCTCCAAACCCAAAAC
djRunx1 7Ro	CAACTTTATATCCAAAAAAC
djRunx1 8Fo	GTTAGGTTTAGGTTTAGA

Continued

Table 2.12: continued from previous page

Name	5'-3' sequence
djRunx1 8Fi	GTTTTGGGTTTGGAGATTAG
djRunx1 8Ri	CACTAAATAAAACAACTC
djRunx1 8Ro	CTTAATTAAAAAACTATCTC
djRunx1 9Fo	TAGAAGTTTTGGGTTTTAAG
djRunx1 9Fi	GAGTTGTTTTATTTAGTG
djRunx1 9Ri	CAAAAAAAAAACCCTTAAAAC
djRunx1 9Ro	AAACACAAACCAACTACTC
djRunx1 10Fo	GTAGGGGTATTTTTTTTATTTTG
djRunx1 10Fi	GAGTTTTAAGGGTTTTTTTTTG
djRunx1 10Ri	CAAACTTCAAACAAAACCTC
djRunx1 10Ro	CTACCTAATCCTCCATAC
djRunx1 11Fo	GAAGAGGAAGAAGTTGTG
djRunx1 11Fi	GAGGTTTTGTTTGAAGTTTG
djRunx1 11Ri	AATTCCCATACAAACTAAAC
djRunx1 11Ro	AACCCTAATTTAAATAAATC
djBSActb1aF	GAAGGTTATAGTTATTTTGG
djBSActb1aR	CTCTCTATCACTAAACAT
djBSActb1bF	TAAGTTTAGGGGATAAAGGA
djBSActb1bR	CCTCTTCCTAATTAACTA
djRunx1P1 1Fo	GTAAAGTTGAGTAAAATTAG
djRunx1P1 1Fi	GTAAAGTTGAGTAAAATTAG
djRunx1P1 1Ri	CAATTTAAATATTTATTTCTAC
djRunx1P1 1Ro	CTAAATACAAACCACAAAC
djRunx1P1 2Fo	GGTATTTTATGGGTTTTATG
djRunx1P1 2Fi	GTAGAAATAAATATTTAAATTG
djRunx1P1 2Ri	ATTAACATCACTTAAATCA
djRunx1P1 2Ro	CAACTATTTCCTTACACAAC
djRunx1P1 3Fo	GTTGTGTAAGGAAATAGTTG
djRunx1P1 3Fi	GAGGAATAATTGATTATTTAG
djRunx1P1 3Ri	CAATAAATTACCAACCTAAC
djRunx1P1 3Ro	TATCATACTATTCCCAAAC

Amplicons were size selected by agarose gel electrophoresis (Invitrogen, 16500-500), and

purified using Zymoclean™ Gel DNA Recovery Kit (Zymo, D4001). 250 ng equimolar PCR amplicons were combined for each biological sample and indexed using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina®. Steps were as described in section 2.8.5.6 except 6 PCR cycles were used.

2.8.7 Illumina® next generation sequencing

Directly prior to sequencing, all next generation sequencing libraries were analysed using a D1000 TapeStation (Agilent), quantified using KAPA Library Quantification Complete kit (Roche, 07960140001), and diluted to 4 nM pooled libraries. Illumina® MiSeq v2 300 cycle kit (Illumina, MS-102-2002), or NextSeq v2 Mid Output 150 cycle kit (TG-160-2001), or NextSeq v2 75 cycle High Output (TG-160-2005) kits were used. WIMM sequencing facility manager Timothy Rostron helped to set up sequencing runs.

2.9 Bioinformatics analysis

Most bioinformatics was performed on a GNU/Linux (2.6.32-754.14.2.el6.x86_64) cluster available through MRC WIMM Centre for Computational Biology. Some analysis including visualisation and statistical testing was performed in RStudio (1.2.1335, R 3.6.0) on my trusty TOSHIBA Satellite running Microsoft Windows 10.0.1734.

2.9.1 Downloading publicly available sequencing data

Publicly available next generation sequencing data generated previously in different cell types were downloaded as fastq files from GEO (Edgar et al. 2002, Table 2.13) using `wget`.

Table 2.13 – GEO accessions of publicly available sequencing data analysed

Cell type	Data type	Accession	Reference
416B	TF / H3K27ac ChIP-seq	GSE69776	Schütte et al. (2016)
416B	DNaseI-seq	GSE37074	Vierstra et al. (2014)
HPC-7	DNaseI-seq, Histone / TF ChIP-seq	E-MTAB-3954	Wilson et al. (2016)
HPC-7	CTCF ChIP-seq	GSE48086	Calero-Nieto et al. (2014)
HPC-7	TF ChIP-seq	GSE22178	Wilson et al. (2010a)
HE/HP	TF / Histone ChIP-seq, DNaseI-seq	GSE69101	Goode et al. (2016)
E14 mESC	CTCF ChIP-seq	GSE22178	Handoko et al. (2011)
E14 mESC	DNaseI-seq	GSE37074	Vierstra et al. (2014)
E14 mESC	Histone ChIP-seq	GSE47950	Wamstad et al. (2012)
V6.5 mESC	Pol II ChIP-seq	GSE20485	Rahl et al. (2010)

2.9.2 ATAC-seq, DNAseI-seq, ChIP-seq data processing

Fastqs were processed using an in-house pipeline (NGseqBasic VS 10.1 or VS 20 Telenius et al. (2018)). Fastq files were trimmed using trim_galore and merging short reads (DNaseI and ATAC-seq data only) using flash (Magoc and Salzberg, 2011).

Mapping to mouse genome mm9 was performed using bowtie with parameters `-p 3 --chunkmb 256 --phred33-quals -m 2 --best --strata --maxins 350 --sam`. Sam files were converted to bam files using samtools. Bam files were filtered using samtools and bedtools. BigWig files were generated from filtered bam files using ucsctools.

2.9.3 bam file processing

Bam files were sorted using `samtools sort` and indexed using `samtools index`.

2.9.4 Peak calling bam files

Peak calling DNaseI-seq data was performed using MACS2/2.0.10 (Zhang et al., 2008) with parameters `-p 0.01 -f BAM -g mm` and using mouse gDNA digested with DNaseI as a control. Number of reads in DNaseI-seq peaks were calculated using an in-house script (Quantbam.pl, J. Hughes, Hay et al. 2016) and peaks were visually inspected using mig in order to fine-tune the threshold number of reads in each peak that gave the best specificity of valid peaks. ChIP-seq data was peak called using MACS2 (Zhang et al., 2008) with parameters `-f BAM -g mm` and using IgG or input tracks as a control where available. Peak sets were visualised using mig and a suitable p value threshold determined between `0.001` and `0.02`. Blacklisted ploidy regions (Yue et al., 2014) with spurious mapping were removed from the peak set using bedtools. A strand-specific TSS database was generated from the refGene curated set of mRNA sequences that were downloaded using UCSC table browser. MACS2 xls peak files were converted to gff files using an in house perl script (macs2gff3.pl, S. Taylor) and then converted to bed files using awk.

2.9.5 Super enhancer annotation

Super enhancers were called using the ROSE algorithm as previously described (Loven et al., 2013; Whyte et al., 2013). 416B DNaseI-seq peaks and signal were used. Awk was used to convert DNaseI peak bed files to ROSE peak files. ROSE_main.py script was run with options `-g mm9 -s 12500 -t 2000`. Awk was then used to convert ROSE_SuperEnhancers.table.txt to a bed file for visualisation in UCSC genome browser.

2.9.6 CTCF motif annotation

A *de novo* CTCF motif was identified in 416B CTCF ChIP-seq data using meme/4.9.1_1 (Bailey and Elkan, 1994). CTCF peaks were called using MACS2 (Zhang et al., 2008) with parameters `-p 0.02 -f BAM -g mm` using 416B Input track as a control. 2000 peaks were sampled using `bedtools sample -n 2000`. Flanking regions for background calculation were calculated from the sampled peaks using `bedtools flank -pct -b 1.00`. Sequences of sampled peaks (samplePeaks.fa) and flanking regions were retrieved from the mm9 genome using `bedtools getfasta`. The background file (CTCF_markov_background.txt) for the model was generated using `fasta-get-markov -m 0`. The *de novo* motif file (Ctcf_meme.txt) was generated using the command `meme samplePeaks.fa -revcomp -dna -nmotifs`

1 -w 20 -maxsize 1000000 -mod zoops -bfile CTCF_markov_background.txt. *De novo* motifs were identified in CTCF peaks or TSS from refGene database using command fimo -motif 1 -bgfile CTCF_markov_background.txt --thresh 1e-3 Ctcf_meme.txt. A single motif with the lowest p value was retained using grep ID=1-1. The output gff file from fimo was converted into bed format using awk.

2.9.7 Capture-C data analysis

Capture-C data was processed essentially as described (Davies et al., 2015) using an in house pipeline developed by Jelena Telenius (Telenius et al., 2018) (CCseqBasic/CM5/pipeRainbow.sh) using parameters -p 4 --genome mm9 --chunkmb 1012 --CCversion CM5 --useSymbolicLinks --BLATforREUSEfolderPath --pf. The pipeline first quality trimmed fastq files using trim_galore, before combining paired-end sequencing reads using flash (Magoc and Salzberg, 2011). Combined reads were then digested *in silico* to DpnII restriction fragments that were mapped using bowtie 2. Mapped reads were only quantified if they contained one captured viewpoint and one reporter fragment with an exclusion size of +/- 1 kb from each viewpoint. PCR duplicates were removed to allow accurate quantification of unique viewpoint-reporter interactions. Gff files resulting from the pipeline were processed manually in R. Interactions per viewpoint were normalised to interactions per 100,000 occurring on the same chromosome (*cis*-normalisation). The mean and standard deviation of three independent biological replicates (n=3) were calculated and exported in bedgraph format. Bedgraphs were visualised as an overlay track in UCSC genome browser using the settings overlay method=solid, windowing function=mean. Quality control on individual replicates was performed in R. Sample clustering was performed on log transformed data using the dist and heatmap functions. PCA analysis was done using plotPCA. Unique *cis*-interactions were calculated by counting the sum of the raw contacts per DpnII fragment per sample and plotted as a box plot using ggplot2. The *cis*-to-*trans* interaction ratio was taken from the COMBINED_allFinalCounts_table.txt generated by the CM5 pipeline. Statistical testing of differential interactions between cell types was performed using DESeq2 (Love et al., 2014) on raw interaction count data using a false discovery rate (FDR) threshold of 0.1. Distance normalisation of interaction data was done on the sum of all three replicates per cell type using peakY (Eijsbouts et al., 2019). Interactions were considered statistically significant using an FDR threshold of 0.1.

2.9.8 Tiled Capture-C data analysis

Tiled Capture-C data was processed in a similar manner to Capture-C data and essentially as described (Oudelaar et al., 2019). For mapping data an in house pipeline developed by Jelena Telenius (CCseqBasic/CM5/pipeRainbow.sh) was used with the following settings -p 8 --pf --BLATforREUSEfolderPath --genome mm9 --chunkmb 1012 --CCversion CM5 --tiled --capturesitesPerBunch 1. COMBINED_CM5_Runx1_mouse.bam files were first processed individually and quality controls performed. Bam files were converted to samfiles using samtools. Sam

files were converted into sparse raw contact matrices using a custom perl script written by Marieke Oudelaar (Oudelaar et al. 2019, Tiled_sam2rawmatrix_MO.pl), using settings `-c 16 -s 91567700 -e 94100269 -b 2000`). ICE normalisation was then done using hicpro (2.11.1) (Servant et al., 2015). ICE normalised matrices were then plotted using an in house python script written by Marieke Oudelaar (using a threshold at the 94th percentile of the data) to check for differences between individual replicates. Individual raw contact matrices were loaded into R and quality control was performed. Sample clustering was performed on log transformed data using the `dist` and `heatmap` functions. PCA analysis was done using `plotPCA`. Unique *cis*-interactions were calculated by counting the sum of the raw contacts per 2 kb bin per sample and plotted as a box plot using ggplot2. The *cis*-to-*trans* interaction ratio was taken from the COMBINED_allFinalCounts_table.txt generated by the CM5 pipeline. To compare Capture-C and Tiled-C interaction profiles, ICE normalised matrices for all six samples were processed into bedgraph files as described in Section 2.9.7 and visualised using UCSC genome browser. After checking that all replicates were similar, sam files for each replicate were merged before being converted to raw sparse contact matrices and ICE normalised together as described above (resulting in one 416B merged sample, and one E14 merged sample). The merged ICE normalised contact matrices were then scaled to 25,000,000 interactions in R to allow meaningful subtractions to be performed. Statistical testing of differential interactions between cell types was performed using DESeq2 (Love et al., 2014) on raw interaction count data from all six independent samples using a false discovery rate (FDR) threshold of 0.1. Normalised matrices, subtraction matrices, and significantly differential contact matrices were then plotted in python as described above.

2.9.9 Bisulfite-seq data processing

Quality of reads in bisulfite sequencing fastq files was checked using fastqc/0.10.1. Reads were quality and adapter trimmed using trim_galore/0.3.1. *In silico* bisulfite converted genome was prepared by Nicki Gray (CBRG) using command
`bismark_genome_preparation --bowtie2 --path_to_bowtie /package/bowtie2/2.1.0/bin --verbose /databank/indices/bismark_bowtie2/mm9`.
Bisulfite sequencing data were mapped using bismark/0.20.0 using the command `bismark --non_directional --bowtie2 --path_to_bowtie /package/bowtie2/2.2.1/bin -q -N 1 /databank/indices/bismark_bowtie2/mm9 -1 read1.fq -2 read2.fq`.
Percentages of methylated CpG dinucleotides were determined using `bismark_methylation_extractor -p --report --merge_non_CpG --bedGraph`. Bedgraph output files were filtered on CpG dinucleotides with coverage greater than 100 reads using `awk -F"\t" '$5+$6>100'`. Coverage bed files were generated using `awk -F"\t" 'print $1"\t"$2"\t"$3'`.

2.9.10 RNA-seq data analysis

Quality of RNA-seq fastq files was checked using fastqc/0.10.1 and trim_galore/0.3.1 was used to trim poor quality reads and adapter sequences. Mapping was performed

using tophat/2.0.13 using the command `tophat --library-type fr-firststrand genome`. Bam files were converted into stranded bigwig files using deeptools/2.2.2 using the command `bamCoverage -bs 250 --normalizeUsingRPKM --filterRNAstrand <forward/reverse>`.

2.9.11 DNaseI footprinting

DNaseI footprinting was performed on publicly available DNaseI-seq data from HE (Goode et al., 2016) and 416B cells (Vierstra et al., 2014) using Sasquatch (Schwessinger et al., 2017). Sasquatch required peak-called and filtered DNaseI-hypersensitive regions to map DNaseI cleaving sites that were identified as the first base in the read from mapped DNaseI-seq bam files. Normalisation was then performed against background (non open chromatin) DNaseI cleavage reads from erythroid cells to account for sequences that are preferentially cut by DNaseI. The `InSilicoMutation` function (Schwessinger et al., 2017) was used to identify damage score hits in *Runx1* enhancer sequences. Relative cut probability was calculated for 7 mer sequences in a 1 bp sliding window over the enhancer sequence. A shoulder to footprint ratio (SFR) value was calculated for each kmer which was then recalculated after each base in the kmer was mutated to all other possible bases. The damage score was calculated by subtracting the SFR of the mutated kmer from the real kmer SFR (Schwessinger et al., 2017). The end result was a single damage score for each base in the enhancer sequence. A damage score greater than 1.0 was considered significant.

2.9.12 Genome-wide CTCF enrichment at transcription start sites

The refGene TSS database was downloaded from UCSC Table Browser and separated into plus and minus strand TSSs using `grep +` and `comm -2 -3`. Bed files containing each TSS with a window of 250 bp, 2000 bp, and 5000 bp (x) centred on each TSS was generated using awk -F'\t' 'print $3, $5-(x/2), $5+(x/2), $13. Non-overlapping TSS (within 5 kb) were isolated using `bedtools merge -c 1`, `awk '/\t1$/print'`, and `bedtools intersect -wa`. CTCF peaks were generated according to Section 2.9.4. Enrichment of CTCF peaks at genomic features and distance to TSS were annotated using homer/4.7 `annotatePeaks.pl mm9` (Heinz et al., 2010) and plotted in R using ggplot. Overlaps of CTCF peaks between samples were performed using homer `mergePeaks` and plotted in R using ggplot. TSS bed files were intersected with peak bed files using `bedtools intersect`. DNaseI, H3K27ac, and CTCF chromatin marks were analysed as in Section 2.9.2. Bam files were converted to bigWig using deeptools `bamCoverage -bs 30 --normalizeUsingRPKM`. Poly A plus bigWig files were generated as in Section 2.9.10 except for using `bamCoverage -bs 30 --normalizeUsingRPKM`. Read counts in a 5 kb bin centred on TSS grouped by distance to TSS were calculated using `bedtools coverage -C`. Read counts for meta-plot heatmaps and profiles were calculated over TSS grouped by distance to TSS using deeptools `computeMatrix scale-regions --missingDataAsZero`. Meta-plot heatmaps were plotted using deeptools `plotHeatmap`. Profiles were plotted in R using data exported from deeptools `plotProfile`. Subset TSS on expression levels was done in R using the `quantile` function. Pairwise correlations between read counts of different chromatin

marks and expression levels were done in R using the `cor` function and plotted using.

2.9.13 Analysis of microhomologies at CRISPR/Cas9 deletions

Large Cas9-induced deletions were sequencing according to Section 2.7.4. Sequences were mapped to the mm9 reference genome using UCSC BLAT (Kent, 2002). A further fine mapping was performed using MUSCLE (Edgar, 2004). Heterozygous deletions were resolved using Poly Peak Parser (Hill et al., 2014). Microhomologies in larger deletions were examined using a custom R package called `mhscanR` (Owens et al., 2019) using the function `mhq`. The function `amh` was used to count the number of alternative microhomologies bypassed during repair of the deletion. The function `gcq` was used to calculate the GC content of microhomologies. The expected background GC content was taken as the average GC content of all the genomic loci containing mapped deletions.

2.9.14 Analysis of deep sequenced short deletions

Fastq files were trimmed using trim_galore. For visualisation, trimmed fastq files were flashed using flash (Magoc and Salzberg, 2011) and mapped using bwa mem v 0.7.12. Sam files were converted to bam using samtools, converted to bigwig using deeptools `bamCoverage -bs 1`, and visualised using the UCSC genome browser. For analysis of individual alleles timed fastq files were analysed using CRISPResso (Pinello et al., 2016). A custom R package called `mhscanR` (Owens et al., 2019) was used to quantify microhomologies in simple deletions (spanning a contiguous region and free of insertions or mutations within 10 bp of the deletion) using the same microhomology scoring criteria as in Section 2.9.13. Reads containing insertions were quantified against the background of all modified reads. The function `gcq` was used to calculate the GC content of microhomologies. The expected background GC content was taken as the average GC content of the region containing 93.5-94.5% (mean +/- two standard deviations) of deletion end points.

2.10 Statistical tests

Statistical testing was done using R. ANOVA was performed using the `aov` function and post hoc testing to perform pairwise comparisons was done using `TukeyHSD`. χ^2 tests were done using the `chisq.test` function. A significance threshold of $p < 0.05$ was used for all statistical tests.

2.10.1 Regression modelling of droplet digital PCR data

Multiple linear regression was performed in R using the `lm` and `predict` functions. The model was fit using the formula $y \sim log(x) + a$, where y (frequency of deletion) was the response variable, x (proximity to sgRNAs) was explanatory variable one and a (sgRNA cutting efficiency) explanatory variable two. Data used to fit the model were empirically determined by ddPCR in eight different targeted cell populations and at two different loci. Frequency of deletion at a particular region was calculated by dividing the relative concentration determined by ddPCR in the transfected sample by the corresponding non-targeting control. Proximity to sgRNA binding sites

was determined from the midpoints of the ddPCR amplicon and sgRNA. The sgRNA cutting efficiency in each sample was the highest frequency of deletion directly at a sgRNA target site as determined by ddPCR. For model estimates of deletion frequency, cutting efficiency was either set to the average cutting efficiency for all samples (when visualising general estimates) or a known cutting efficiency when estimating values in a subset of the experiments. Goodness of fit testing was performed by plotting a histogram of residuals (where $residual = observed - predicted$), and a Q-Q plot (quantile-quantile plot) to check that residuals are normally distributed (which is an assumption of the linear regression model). The distribution of estimated relative allele frequencies ($1 - deletion frequency$) was plotted across a 3 kb window up and downstream of a simulated sgRNA cut site.

3. Characterising *cis*-interactions in the transcriptionally active and inactive *Runx1* regulatory domain

3.1 Introduction

The *Runx1* gene is large and complex (Figure 1.3). It spans 224kb on chromosome 16 in mouse and 21 in human, and has two alternative promoters (P1 and P2) both with distinct spatiotemporal activities (Miyoshi et al., 1995; Ghozi et al., 1996; Telfer and Rothenberg, 2001; Pozner et al., 2007; Sroczynska et al., 2009a; Bee et al., 2009b, 2010). However, the two promoters do not confer haematopoietic-specific expression (Ghozi et al., 1996; Bee et al., 2009b), indicative of its tissue-specific regulation being mainly orchestrated by enhancers. Several different *Runx1* enhancers were identified in our laboratory based on multi-species sequence conservation, DNaseI hypersensitivity, haematopoietic TF binding, histone acetylation, and transgenic enhancer-reporter assays *in vitro* and *in vivo* (Nottingham et al., 2007; Schütte et al., 2016). Several of these enhancers confer haematopoietic-specific expression haematopoietic sites in a *Runx1*-specific pattern (Figure 1.4, Nottingham et al. 2007; Bee et al. 2009b, 2010; Swiers et al. 2013a; Schütte et al. 2016). It remains to be seen, however, whether the previously identified haematopoietic enhancers interact with the *Runx1* promoters in haematopoietic cells, and what drives the changes in *cis*-interactions to facilitate enhancer-promoter interactions.

Chromosome conformation capture (3C) is a powerful technique for coupling enhancers to the promoter(s) they regulate (Dekker et al., 2002; Tolhuis et al., 2002). NG Capture-C (hereafter referred to as Capture-C) allows high-resolution analysis of chromatin interactions at the restriction fragment level at selected sites such as gene promoters, enhancers, and CTCF sites in an one-vs-all experiment (Hughes et al., 2014; Hay et al., 2016; Hanssen et al., 2017; Oudelaar and Higgs, 2017). Tiled-C combines the high-resolution of Capture-C with the all-vs-all unbiased nature of Hi-C over an entire region of interest up to several Mb (Oudelaar et al., 2019).

In order to examine differences in *cis*-interactions that can be observed between cells where *Runx1* is transcriptionally inactive and active, I chose two model cell lines. The first is a murine haematopoietic myeloid progenitor cell line (416B, Dexter et al. 1979) that express high levels of P1 and P2-derived *Runx1* transcripts (Ben-Ami et al., 2009). The second are E14 mouse embryonic stem cells (mESCs) that acted as a control cell type due to their pluripotent embryonic state and lack of lineage-specific transcription factor (TF) expression (Handyside et al., 1989). To examine *cis*-interactions within the *Runx1* domain in both cell types, I performed Capture-C using *Runx1* promoters and enhancers that likely play important roles in regulating transcriptional activity in the domain. Interactions from the viewpoint of neighbouring gene promoters and boundary CTCF sites were also determined in order to

delimit the extent of the *Runx1* regulatory domain. Finally, the overall structure of the *Runx1* domain was elucidated using Tiled-C.

3.2 Results

3.2.1 Quality control of Capture-C data

For Capture-C analysis of the *Runx1* locus, three independent biological replicates of each of the two cell lines used–E14 mESCs and 416B haematopoietic progenitors–were used. Unsupervised hierarchical clustering and principal component analysis (PCA) of unique *cis*-interaction (reporter) counts revealed that the biological replicates clustered together well, while the two different cell types were grouped distinctly (Figure 3.1 a, b). In total, 2,220,083 and 1,703,962 reporters were quantified in 416B and E14 cells, respectively. Between all six samples, the number of reporters per viewpoint ranged from 8957 to 85785 over all fifteen viewpoints (Figure 3.1 c). Viewpoints included the two *Runx1* promoters (P1 and P2), *Runx1* enhancers (+23, +110, +204), CTCF binding sites (-718, -629, -468, +260, +314, +380), and neighbouring gene promoters (*Cbr1*, *Dopey2*, *Clic6*, and *Rcan1*). The percentage of *cis* to *trans* interactions (which is a proxy for enrichment) ranged from 58.4% to 85.5% (Figure 3.1 d). All these quality controls indicate the high quality of these data.

3.2.2 Defining the *cis*-regulatory interactions from the viewpoint of both *Runx1* promoters

Visual inspection of Capture-C interaction profiles generated for the *Runx1* promoters P1 and P2 showed that they interacted widely within a large ∼1.1 megabase (Mb) regulatory domain (Figure 3.2 a). Mean reporter count-normalised interaction profiles in 416B cells and E14 mESCs showed that the promoters interacted with the entire gene desert adjacent to the gene (Figure 3.2 b, c). Interestingly, this large interaction domain roughly agrees with the evolutionarily conserved (human-mouse-frog) synteny block identified previously (Ahituv et al., 2005) (Figure 3.2 a). However, it also extends further in the centromeric direction towards the genes *Clic6* and *Rcan1* within the paralogous region (Figure 3.2 b, c). Distance-normalisation was performed to account for the 'proximity signal' in the Capture-C data–which was evident as a peak in interactions that decayed exponentially with linear distance away from the chosen viewpoint. Statistical testing of distance-normalised interaction data identified significant long-range interactions between the promoters and regions in the gene desert in both cell types (Figure 3.2, dist.norm. tracks in b and c, *, FDR < 0.1).

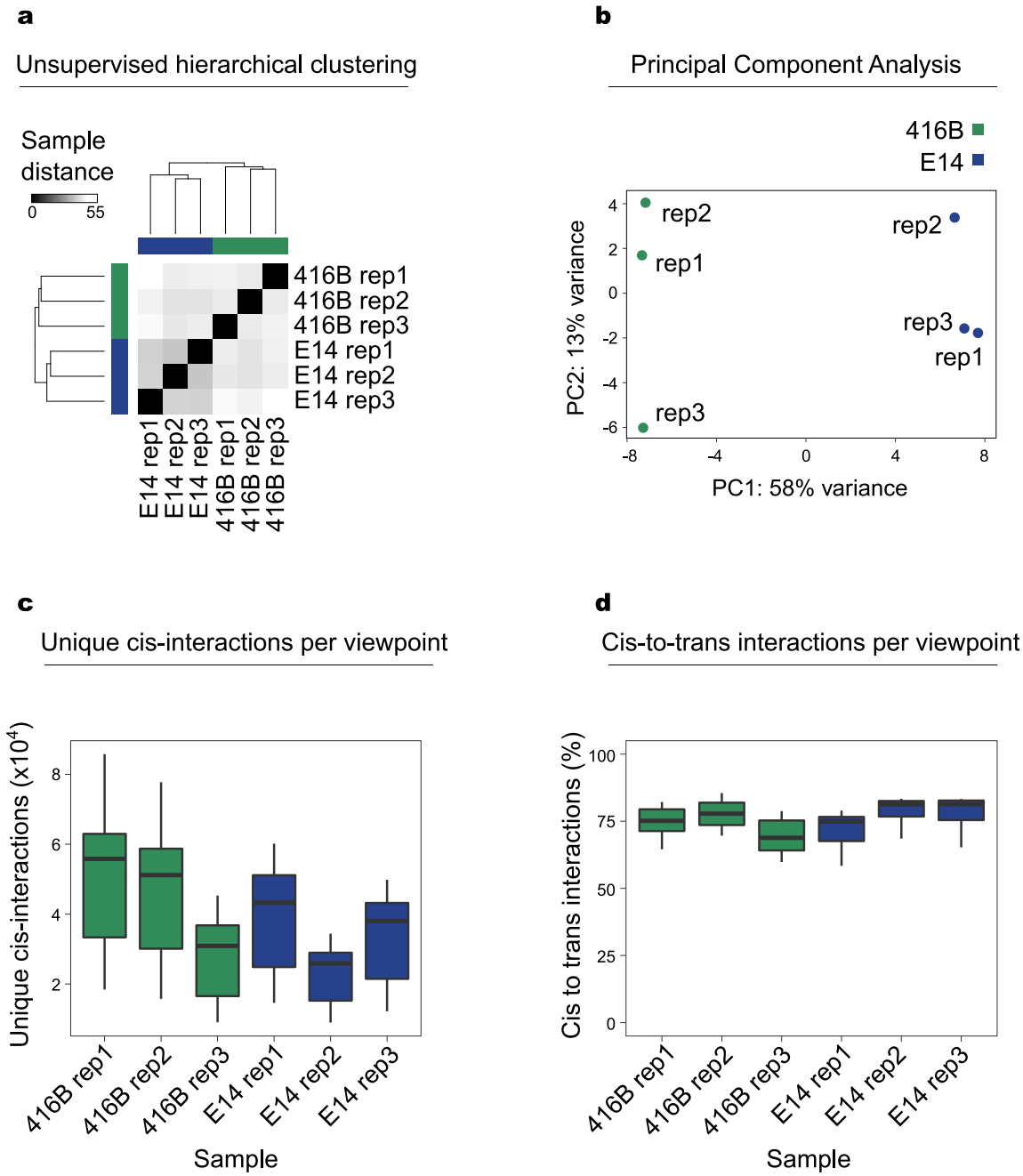


Figure 3.1 – Quality control of Next Generation Capture-C data.
a) Sample distances and unsupervised hierarchical clustering of six independent Capture-C samples. b) Principle component analysis of six independent Capture-C samples. c) Quantification of unique PCR duplicate filtered *cis*-interactions per DpnII restriction fragment in each sample. d) *Cis*-to-*trans* interaction quantification for each viewpoint in each sample.

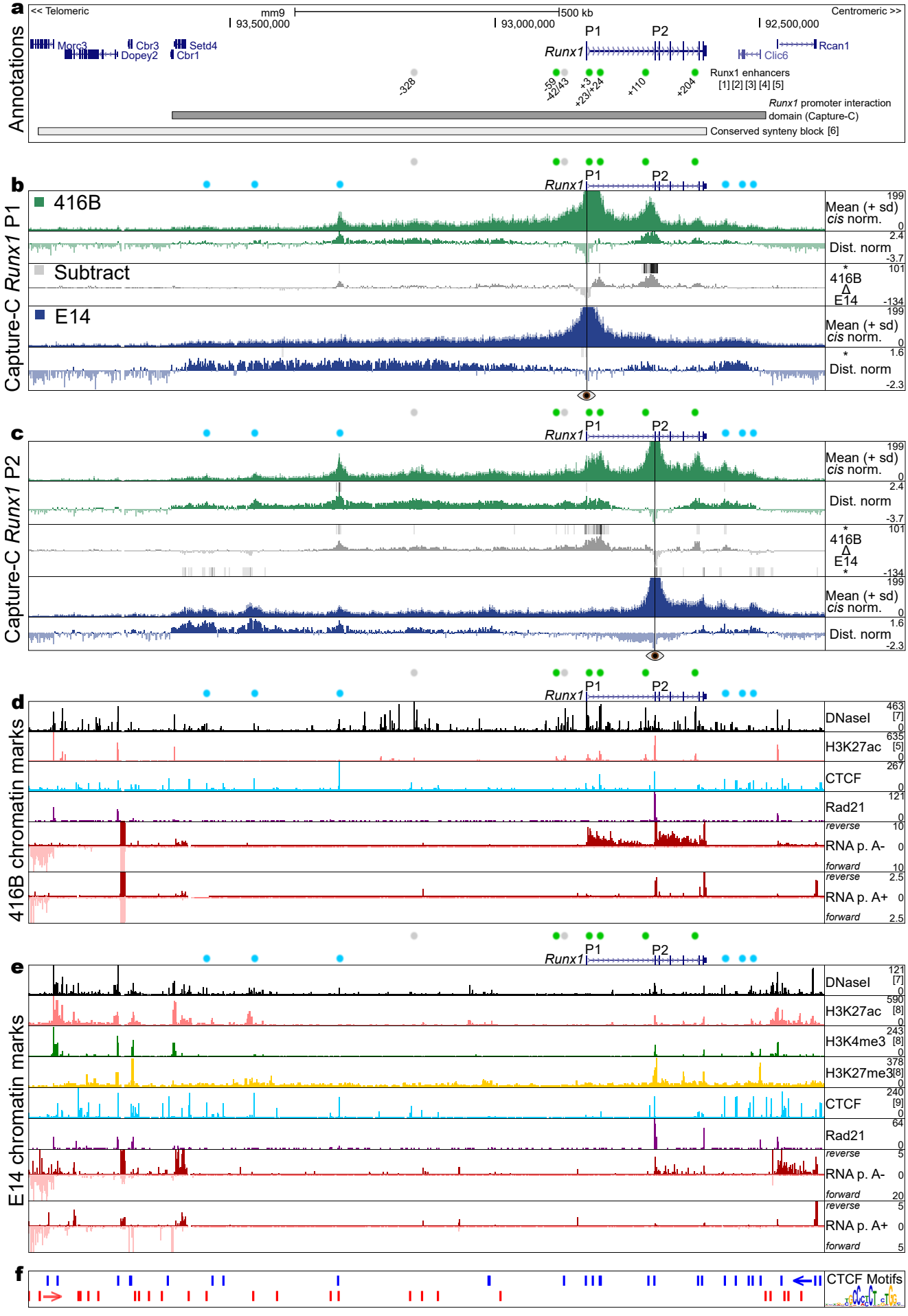


Figure 3.2 – Capture-C from the viewpoint of *Runx1* promoters. *Legend continued on next page.*

Figure 3.2 – *Legend continued from previous page.* a) Annotations showing coding genes (UCSC), the locations of *Runx1* enhancer elements previously identified ([1] Nottingham et al. 2007, [2] Bee et al. 2009b, [3] Bee et al. 2010, [4] Swiers et al. 2013a, [5] Schütte et al. 2016) and shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles). The span of an evolutionarily conserved block of synteny previously identified is shown ([6] Ahituv et al. 2005), and the approximate extent of the *Runx1* promoters interaction domain defined here by Capture-C is shown for comparison. Blue circles indicate regions that interacted with the *Runx1* promoters and were associated with CTCF binding sites. b-c) Capture-C from the viewpoint of both *Runx1* promoters in 416B haematopoietic cells (green tracks) and E14 mESCs (blue tracks). The mean reporter counts from three independent samples (n=3) normalised to the total *cis*-interactions in each sample are shown in the dark colour, with the mean plus standard deviation shown in the lighter colour above the 'Mean (+ sd) *cis* norm.' tracks. Distance-normalised tracks are shown, with interacting fragments significantly different (FDR < 0.1) from the expectation based on linear distance highlighted as bars above or below the Dist. norm tracks. A subtraction of signal in E14 cells from 416B cells is shown in grey (416B Δ E14) with interacting fragments significantly different (FDR < 0.1) between the two cell types highlighted as bars above or below the subtraction tracks. The vertical line marked with an eye icon indicates the position of each viewpoint. d-e) Chromatin marks in 416B cells and E14 mESCs either downloaded from NCBI GEO and analysed by me or generated and analysed by me. Public data analysed were 416B and E14 DNaseI-seq ([7] Vierstra et al. 2014), 416B H3K27ac ([5] Schütte et al. 2016), E14 H3K27ac, H3K4me3, H3K27me3 ([8] Wamstad et al. 2012), E14 CTCF ([9] Handoko et al. 2011). f) CTCF peaks called in 416B cells with *de novo* motif discovery. Blue and red bars indicate CTCF motifs oriented in the telomeric and centromeric direction, respectively. UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_CapC_proms`

Statistical testing between the two cell types identified regions that interacted significantly more frequently in 416B cells compared to E14 mESCs (Figure 3.2, bars atop the 416B Δ E14 tracks in b and c, *, FDR < 0.1). Annotating the positions of the *Runx1* enhancers previously identified in our laboratory (-328, -42/-43, -59, +23, +110, +204) revealed that all these regions interacted significantly more frequently with either P1 or P2 promoter in 416B cells compared to E14 mESCs (Figure 3.2 a - d). In 416B cells, the P1 promoter interacted strongly with the +23 and +110 enhancers, but less so with the +204 enhancer. Unexpectedly, some regions also interacted more frequently in E14 cells compared to 416B cells (Figure 3.2 bars below the 416B Δ E14 tracks in b and c, *, FDR < 0.1). Specifically, the P2 promoter interacted with the far end of the gene desert, and the region containing the genes *Clic6* and *Rcan1*, more frequently in E14 mESCs than 416B cells.

To examine transcriptional activity within the *Runx1* regulatory domain defined by Capture-C, I curated a set of chromatin epigenetic marks generated in both cell types. The *Runx1* enhancers previously identified (Nottingham et al., 2007; Schütte et al., 2016) were enriched for DNaseI-seq hypersensitivity and H3K27ac ChIP-seq signal in 416B cells but not in E14 mESCs (Figure 3.2, d and e). Poly-A+ and poly-A- RNA-seq, quantifying nascent and mature transcripts, respectively, showed that 416B cells were transcribing *Runx1* from both P1 and P2 promoters at the population level (Figure 3.2 d). Unexpectedly, nascent and mature transcripts originating from the P2 promoter were also seen in E14 mESCs (Figure 3.2 e). Publicly available ChIP-seq in E14 mESCs showed that the P2 promoter was marked with both active (H3K4me3, H3K27ac) and repressive (H3K27me3) histone modifications, which is characteristic of a bivalent promoter (Bernstein et al., 2006). Transcript levels were notably higher in 416B cells compared to mESCs, however, and enhancer-promoter interactions occurred significantly more frequently (Figure 3.2 a and b). Capture-C therefore, provides strong evidence that the *Runx1* enhancer elements previously identified in our laboratory are capable of interacting with *Runx1* promoters to drive *Runx1* transcription in a haematopoietic progenitor cell line.

3.2.3 The *Runx1* regulatory domain is bounded by clusters of convergently oriented CTCF binding sites

Capture-C also revealed several regions that were not associated with active transcription marks and interacted strongly with the two *Runx1* promoters (Figure 3.2 b and c, blue circles). A candidate for driving these interactions is CTCF/cohesin. To investigate this hypothesis, I performed CTCF and Rad21 (a component of the cohesin complex) ChIP-seq in 416B cells and Rad21 ChIP-seq in E14 mESCs (along with analysis of publicly available CTCF ChIP-seq generated previously in E14 mESCs Handoko et al. 2011). CTCF binding was observed at many of the non-enhancer regions that interacted with the *Runx1* promoters, while Rad21 binding was restricted mainly to the P2 promoter in both cell types (Figure 3.2, d and e). CTCF was also bound at the P2 promoter in both 416B and E14 mESCs, while binding at the P1 promoter was most prominent in 416B cells (Figure 3.2, d and e). Multiple CTCF binding sites were observed at the boundaries of the *Runx1* regulatory domain (Figure 3.2, d and e). Peak calling and *De novo* motif discovery revealed typical CTCF binding motifs under these peaks. Analysis of motif orientation revealed that the two

clusters of CTCF binding sites at the boundaries of the *Runx1* regulatory domain contained several motifs in the same orientation (Figure 3.2, f). When looking at both clusters together, they generally seem to occur in a convergent orientation to each other (Figure 3.2, f, generally 'red' oriented motifs at the telomeric boundary of the domain and exclusively 'blue' oriented at the centromeric boundary).

To shed further light on the extent of the *Runx1* interacting domain I visualised Capture-C interaction profiles from the boundary CTCF sites (Figures 3.3 and 3.4). Generally, interaction frequencies from CTCF sites towards the edges (outermost) extent of the domain (-718, +380) interacted little with the *Runx1* gene, and tended to interact with the cluster of CTCF sites at the other end of the domain (Figures 3.3 b and 3.4 d). Interactions between the outermost CTCF sites were tissue-specific and occurred significantly more frequently in 416B cells than in E14 mESCs (Figures 3.3 b and c, 3.4 c and d), *, FDR < 0.1). In contrast, CTCF sites more proximal to the *Runx1* gene (-468, +260) interacted frequently within the *Runx1* domain (Figures 3.3 d and 3.4 b). Interestingly, the -468 CTCF site in the middle of the gene desert interacted frequently with both *Runx1* gene promoters (Figure 3.3 d, *, FDR < 0.1), and these interactions occurred significantly more frequently in 416B cells compared to E14 cells. Distance-normalisation revealed that the -468 CTCF site interacted with the entire *Runx1* domain above the expected interaction frequency based on linear distance, with peaks of interaction significantly greater than expected at the promoters (Figure 3.3 d, *, FDR < 0.1). Collectively, these analyses identify specific convergently oriented CTCF binding sites that form the boundaries of the *Runx1* regulatory domain.

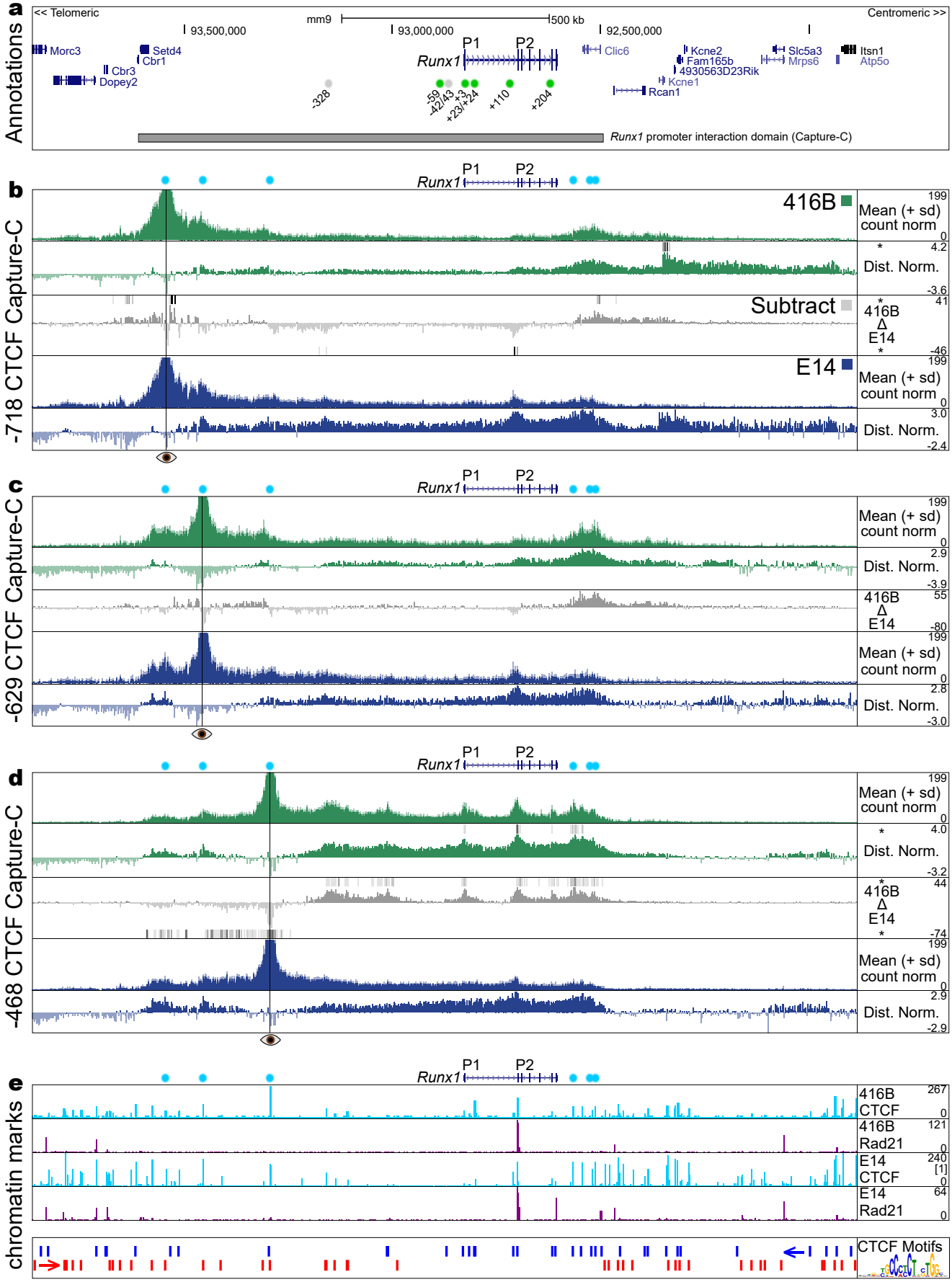


Figure 3.3 – Capture-C from the viewpoint of CTCF sites in the gene desert adjacent to *Runx1*. *Legend continued on next page.*

Figure 3.3 – *Legend continued from previous page.* a) Annotations showing coding genes (UCSC), the locations of *Runx1* enhancer elements previously identified (Nottingham et al., 2007; Bee et al., 2009b, 2010; Swiers et al., 2013a; Schütte et al., 2016) and shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles). The extent of the *Runx1* promoters interaction domain defined by Capture-C is shown. Blue circles indicate regions that interacted with the *Runx1* promoters and were associated with CTCF binding sites. b-d) Capture-C from the viewpoint of -718, -629, and -468 CTCF binding sites in 416B haematopoietic cells (green tracks) and E14 mESCs (blue tracks). The mean reporter counts from three independent samples (n=3) normalised to the total *cis*-interactions in each sample are shown in the dark colour, with the mean plus standard deviation shown in the lighter colour above the 'Mean (+ sd) *cis* norm.' tracks. Distance-normalised tracks are shown, with interacting fragments significantly different (FDR < 0.1) from the expectation based on linear distance highlighted as bars above or below the Dist. norm tracks. A subtraction of signal in E14 cells from 416B cells is shown in grey (416B Δ E14) with interacting fragments significantly different (FDR < 0.1) between the two cell types highlighted as bars above or below the subtraction tracks. The vertical line marked with an eye icon indicates the position of each viewpoint. e) Chromatin marks in 416B and E14 cells. Public data analysed were E14 CTCF ChIP-seq ([1] Handoko et al. 2011). CTCF peaks called in 416B cells with *de novo* motif discovery. Blue and red bars indicate CTCF motifs oriented in the telomeric and centromeric direction, respectively. UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_CapC_Ctcf_left`

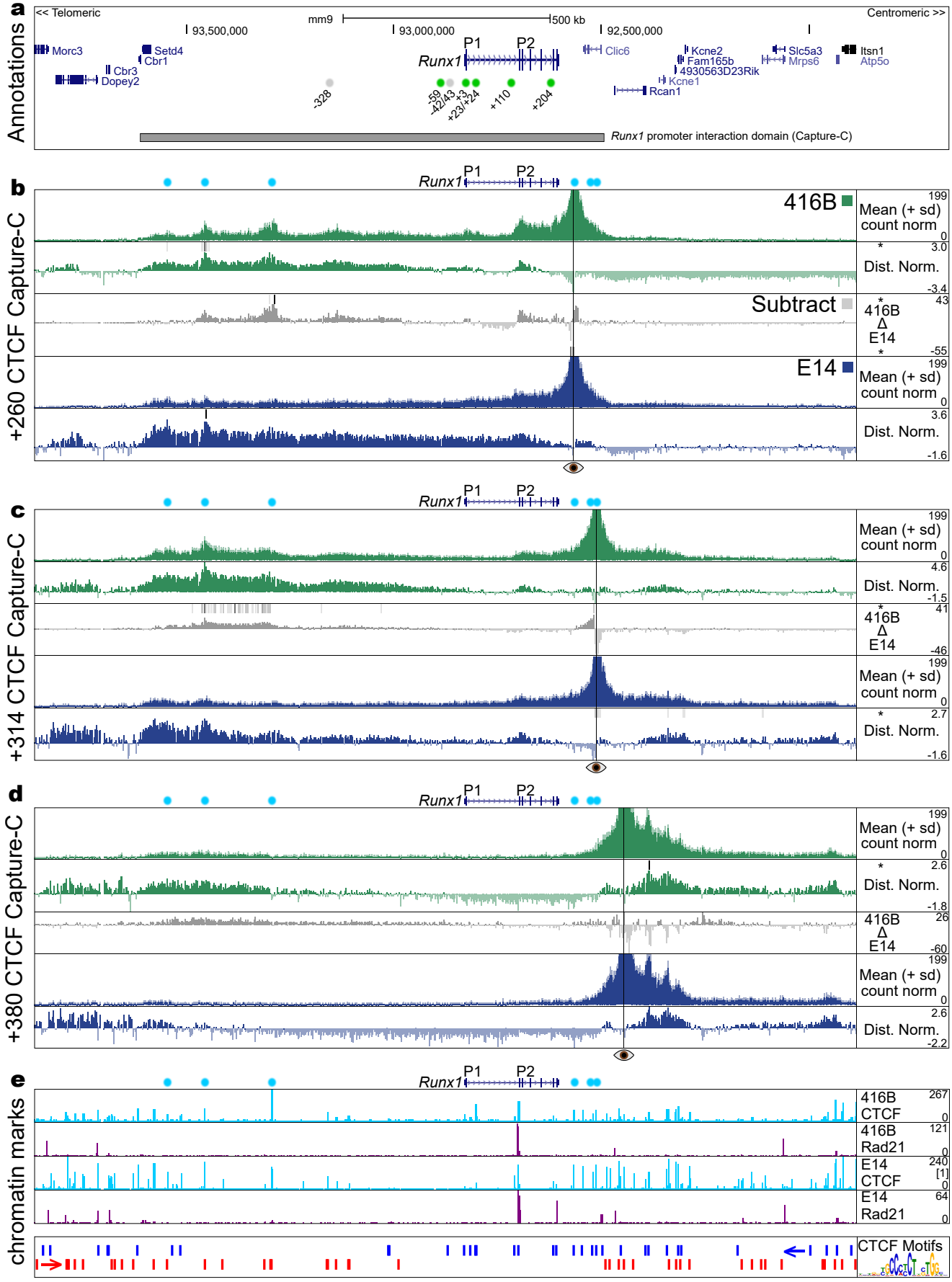


Figure 3.4 – Capture-C from the viewpoint of CTCF sites in the region centromeric to *Runx1* containing the genes *Clic6* and *Rcan1*. *Legend continued on next page.*

Figure 3.4 – *Legend continued from previous page.* a) Annotations showing coding genes (UCSC), the locations of *Runx1* enhancer elements previously identified (Nottingham et al., 2007; Bee et al., 2009b, 2010; Swiers et al., 2013a; Schütte et al., 2016) and shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles). The extent of the *Runx1* promoters interaction domain defined by Capture-C is shown. Blue circles indicate regions that interacted with the *Runx1* promoters and were associated with CTCF binding sites. b-d) Capture-C from the viewpoint of +260, +314, and +380 CTCF binding sites in 416B haematopoietic cells (green tracks) and E14 mESCs (blue tracks). The mean reporter counts from three independent samples (n=3) normalised to the total *cis*-interactions in each sample are shown in the dark colour, with the mean plus standard deviation shown in the lighter colour above the 'Mean (+ sd) *cis* norm.' tracks. Distance-normalised tracks are shown, with interacting fragments significantly different (FDR < 0.1) from the expectation based on linear distance highlighted as bars above or below the Dist. norm tracks. A subtraction of signal in E14 cells from 416B cells is shown in grey (416B Δ E14) with interacting fragments significantly different (FDR < 0.1) between the two cell types highlighted as bars above or below the subtraction tracks. The vertical line marked with an eye icon indicates the position of each viewpoint. e) Chromatin marks in 416B and E14 cells. Public data analysed were E14 CTCF ChIP-seq ([1] Handoko et al. 2011). CTCF peaks called in 416B cells with *de novo* motif discovery. Blue and red bars indicate CTCF motifs oriented in the telomeric and centromeric direction, respectively. UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_CapC_Ctcf_right`

3.2.4 Enhancer interactions in the *Runx1* domain and beyond

Since *Runx1* resides within an evolutionarily conserved synteny block (Ahituv et al., 2005), we wondered whether cis-regulatory mechanisms between *Runx1* and its neighbouring genes may be shared. For example, it is plausible that enhancers within the *Runx1* domain also interacted with and affected neighbouring genes, or vice versa. To address this question, I first plotted Capture-C from the viewpoint of the *Runx1* enhancers +23, +110, and +204 (Figure 3.5). In agreement with Capture-C profiles from the viewpoint of *Runx1* promoters, the enhancers interacted more with the promoters in 416B cells than E14 mESCs (Figures 3.5 b, c, and d, *, FDR $<$ 0.1). Capture-C from the *Runx1* enhancers also revealed that they interacted only infrequently with genes neighbouring *Runx1* (Figure 3.5 b, c, and d). Distance-normalisation revealed that the enhancers generally interacted less frequently than would be expected given the relatively short linear distances between them and the two nearest genes *Clic6* and *Rcan1*. However, in E14 mESCs substantial (but not significant) interactions between *Runx1* enhancers and the body of *Clic6* did occur (Figure 3.5 b, c, and d). A similar interaction profile in mESCs was also seen for the Runx1 P2 promoter (Figure 3.2 b and c). RNA-seq showed that *Runx1* and *Clic6* expression was minimal in both cell types, suggesting that these interactions were distinct from typical enhancer-promoter interactions involved in transcriptional regulation of the genes.

Interestingly, weak but significant long-range interactions were detected between *Runx1* enhancers and an alternative promoter of the gene *Erg*, located around 3 Mb telomeric to *Runx1* (Figures 3.5 b, and c, *, FDR $<$ 0.1). This interaction was only seen in 416B cells and was significant for +110 (and observable for +23, +204 and the *Runx1* promoters [promoters not shown]). Both *Runx1* and *Erg* were highly expressed in 416B cells by RNA-seq (Figure 3.5 e). An analysis to rank super-enhancers (SEs, by stitching together DNaseI-seq peaks and assigning a DNaseI-seq read depth score) in 416B cells, identified both the *Runx1* +23 and *Erg* alternative promoter as putative SEs (Figure 3.5 a, 7.1). Even though this region of *Erg* is in fact an alternative promoter and likely not a super-enhancer, this suggests that *Runx1* enhancers engage in long-distance interactions, plausibly mediated by phase separation of transcriptionally active regions.

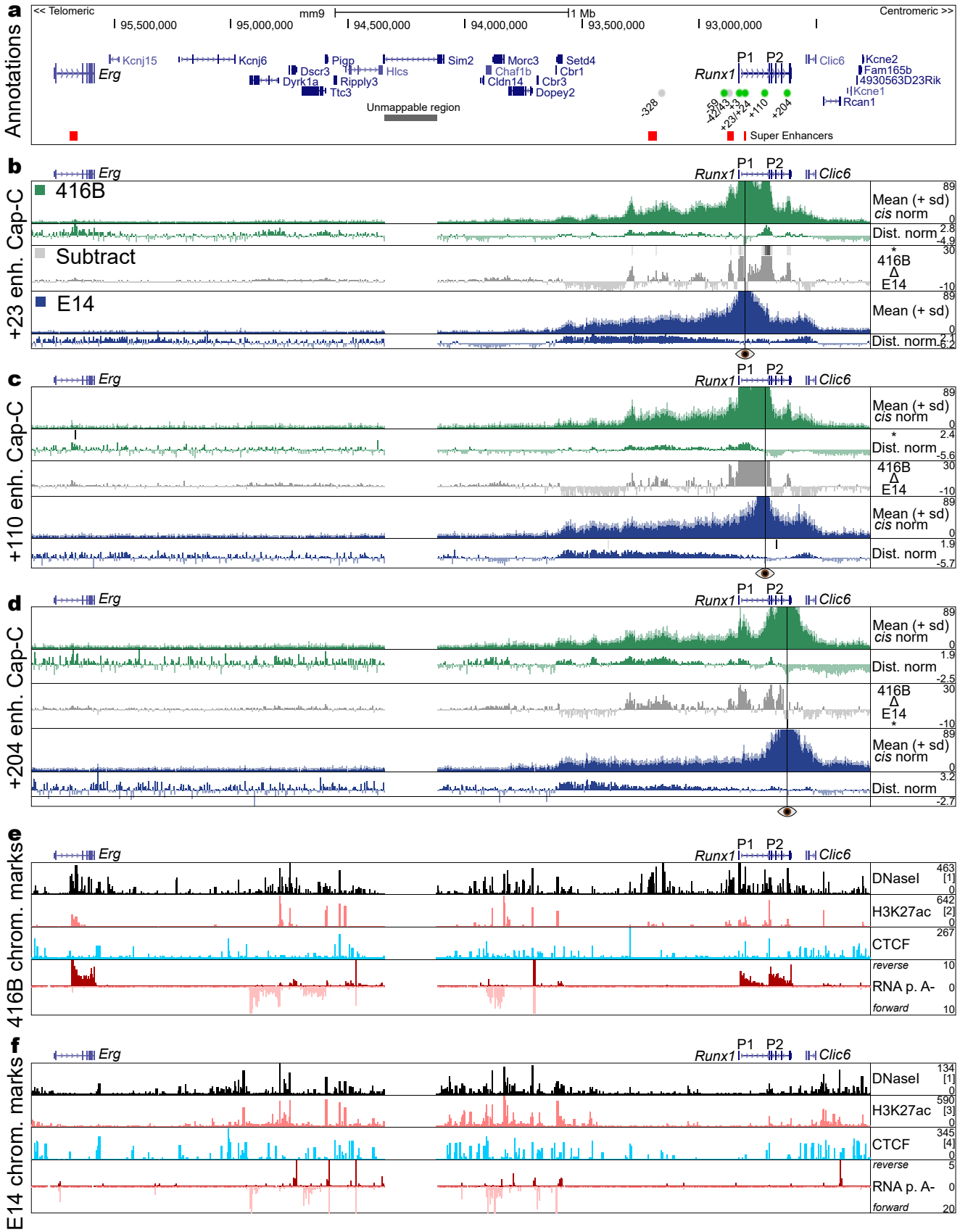


Figure 3.5 – Capture-C from the viewpoint of *Runx1* enhancers. *Legend continued on next page.*

Figure 3.5 – *Legend continued from previous page.* a) Annotations showing coding genes (UCSC), the locations of *Runx1* enhancer elements previously identified (Nottingham et al., 2007; Bee et al., 2009b, 2010; Swiers et al., 2013a; Schütte et al., 2016) and shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles). Super-enhancers identified in 416B cells are shown as red boxes. b-d) Capture-C from the viewpoint of +23, +110, and +204 enhancers in 416B haematopoietic cells (green tracks) and E14 mESCs (blue tracks). The mean reporter counts from three independent samples (n=3) normalised to the total *cis*-interactions in each sample are shown in the dark colour, with the mean plus standard deviation shown in the lighter colour above the 'Mean (+ sd) *cis* norm.' tracks. Distance-normalised tracks are shown, with interacting fragments significantly different (FDR < 0.1) from the expectation based on linear distance highlighted as bars above or below the Dist. norm tracks. A subtraction of signal in E14 cells from 416B cells is shown in grey (416B Δ E14) with interacting fragments significantly different (FDR < 0.1) between the two cell types highlighted as bars above or below the subtraction tracks. The vertical line marked with an eye icon indicates the position of each viewpoint. e-f) Chromatin marks in 416B and E14 cells. Public data analysed were 416B and E14 DNaseI-seq ([1] Vierstra et al. 2014), 416B H3K27ac ([2] Schütte et al. 2016) E14 H3K27ac ([3] Wamstad et al. 2012) E14 CTCF ([4] Handoko et al. 2011). UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_CapC_enhancers_Erg_wide`

Next, I analysed Capture-C profiles from the viewpoint of the promoters of four genes neighbouring *Runx1–Cbr1*, *Dopey2*, *Clic6*, and *Rcan1* (Figure 3.6). In agreement with Capture-C from the *Runx1* enhancers, these neighbouring gene promoters were generally excluded from the *Runx1* domain. Distance-normalisation revealed that the genes neighbouring *Runx1* generally interacted with the *Runx1* domain less than would be expected based on the linear distance between them (Figure 3.6 b-e). *Clic6* was the only gene promoter that interacted significantly with the *Runx1* domain out of all the neighbouring gene promoters analysed (Figure 3.6 d). The *Clic6* promoter had a broad region of significant interactions in 416B cells around the cluster of CTCF binding sites that make up the boundary within the gene desert. *Clic6* also interacted with the P2 promoter in E14 mESCs, and, interestingly, both were bound by Pol II (Figure 3.6 Pol II track in g; reanalysis of public Pol II ChIP-seq Rahl et al. 2010). However, in 416B cells when *Runx1* is highly expressed, none of these neighbouring gene promoters interacted significantly with the *Runx1* promoters or enhancers.

3.2.5 Structure within the *Runx1* domain is dramatically altered when transcription is active

3.2.5.1 Quality control of Tiled-C data

Due to the complexity and tissue-specificity of the *cis*-interactions mapped from the various viewpoints by Capture-C it became important to understand the overall structure of the region and observe how this changes with transcriptional activation. Therefore, in order to gain a comprehensive view of the entire repertoire of *cis*-interactions within the *Runx1* domain that are associated with high and low levels of transcription, I performed Tiled-C (Oudelaar et al., 2019). Instead of visualising interactions from a small number of DpnII restriction fragments as in Capture-C, Tiled-C utilised all 4397 DpnII fragments as viewpoints within 2.5 Mb centred on *Runx1*. In order to facilitate visualisation of these data (as an all-vs-all Hi-C-like contact matrix), it was necessary to bin the DpnII fragments into 2 kb bins. After sequencing, 46,159,582 and 37,139,816 unique interactions between bins (reporters) were recovered in 416B and E14 cells, respectively (n=3). The number of reporters per bin ranged from 0 to 47017 between all six samples (mean=10948, sd=6354), and the percentage of *cis* to *trans* interactions ranged from 93.8% to 96.0% (Figure 3.7 c, d). Clustering raw reporter counts revealed that both cell types clustered distinctly while biological replicates clustered together (Figure 3.7 a, b). Moreover, comparison between Capture-C and Tiled-C interaction profiles for *Runx1* P1 and P2 promoters revealed that the quality of the data was comparable, even though the Tiled-C experiment involved thousands of viewpoints analysed at once (Figure 3.7 e). Comparison between Tiled-C data at *Runx1* in E14 mESCs and the highest-resolution Hi-C data available (Bonev et al., 2017) highlighted the superior depth of the Tiled-C data (Figure 3.8).

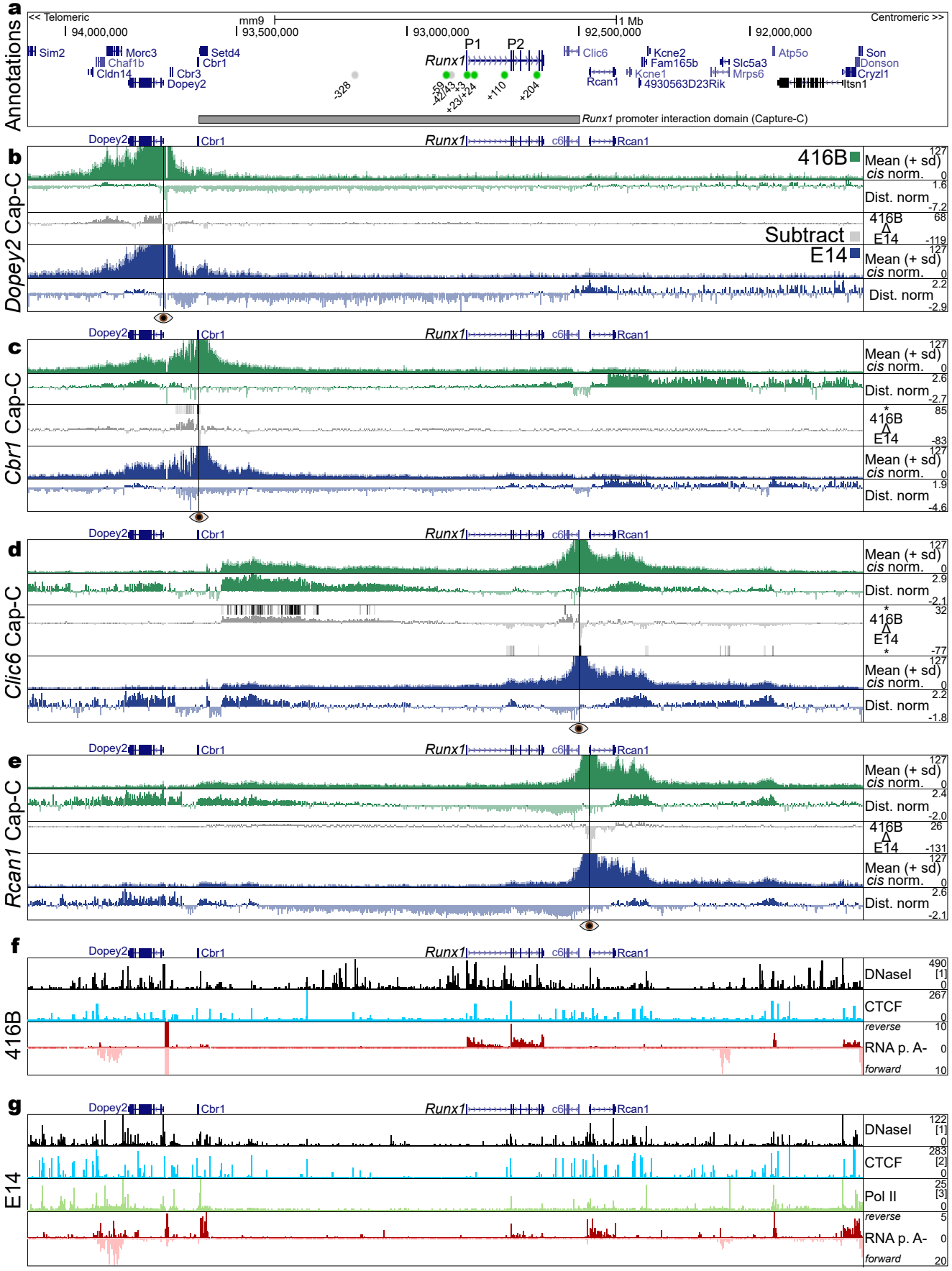


Figure 3.6 – Capture-C from the viewpoint of gene promoters neighbouring *Runx1*.
Legend continued on next page.

Figure 3.6 – *Legend continued from previous page.* a) Annotations showing coding genes (UCSC), the locations of *Runx1* enhancer elements previously identified (Nottingham et al., 2007; Bee et al., 2009b, 2010; Swiers et al., 2013a; Schütte et al., 2016) and shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles). The extent of the *Runx1* promoters interaction domain defined by Capture-C is shown. b-e) Capture-C from the viewpoint of the promoters of *Dopey2*, *Cbr1*, *Clic6*, *Rcan1* in 416B haematopoietic cells (green tracks) and E14 mESCs (blue tracks). The mean reporter counts from three independent samples (n=3) normalised to the total *cis*-interactions in each sample are shown in the dark colour, with the mean plus standard deviation shown in the lighter colour above the 'Mean (+ sd) *cis* norm.' tracks. Distance-normalised tracks are shown (Dist. norm tracks). A subtraction of signal in E14 cells from 416B cells is shown in grey (416B Δ E14) with interacting fragments significantly different (FDR < 0.1) between the two cell types highlighted as bars above or below the subtraction tracks. The vertical line marked with an eye icon indicates the position of each viewpoint. f-g) Chromatin marks in 416B and E14 cells. Public data analysed were 416B and E14 DNaseI-seq ([1] Vierstra et al. 2014), E14 CTCF ([2] Handoko et al. 2011), E14 RNA polymerase II (Pol II, [3], Rahl et al. 2010). UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_CapC_neighbourGenes_both`

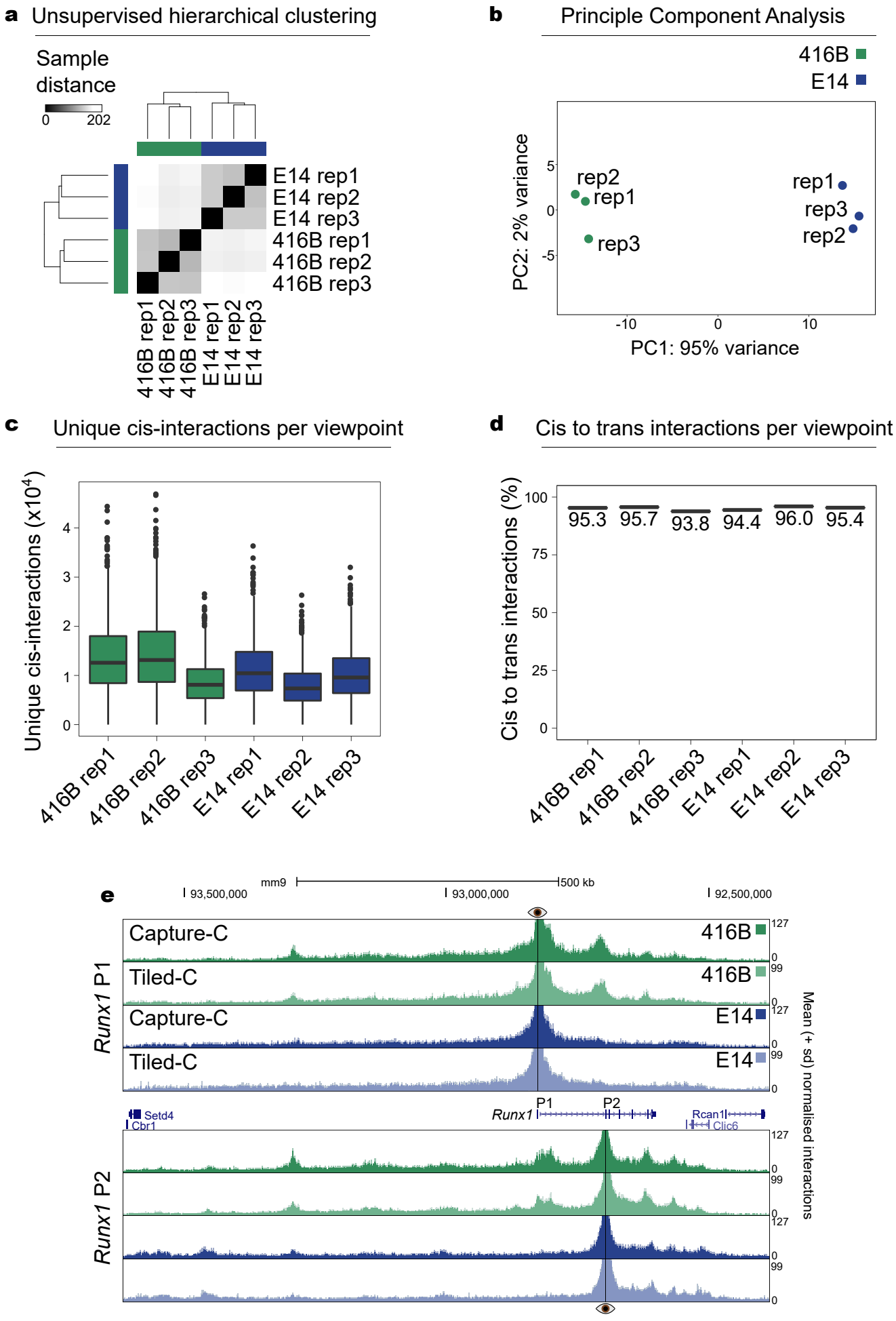


Figure 3.7 – Quality control of Tiled-C data. *Legend continued on next page.*

Figure 3.7 – *Legend continued from previous page.* a) Sample distances and unsupervised hierarchical clustering of six independent Tiled-C samples. b) Principle component analysis of six independent Tiled-C samples. c) Quantification of unique PCR duplicate filtered *cis*-interactions per 2 kb bin of DpnII restrict fragments in each sample. d) *Cis*-to-*trans* interaction quantification for each bin in each sample. e) Comparison between Capture-C and Tiled-C interaction profiles for the *Runx1* promoters in 416B and E14 cells. UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_CapC_proms_TiledvirtCapC`

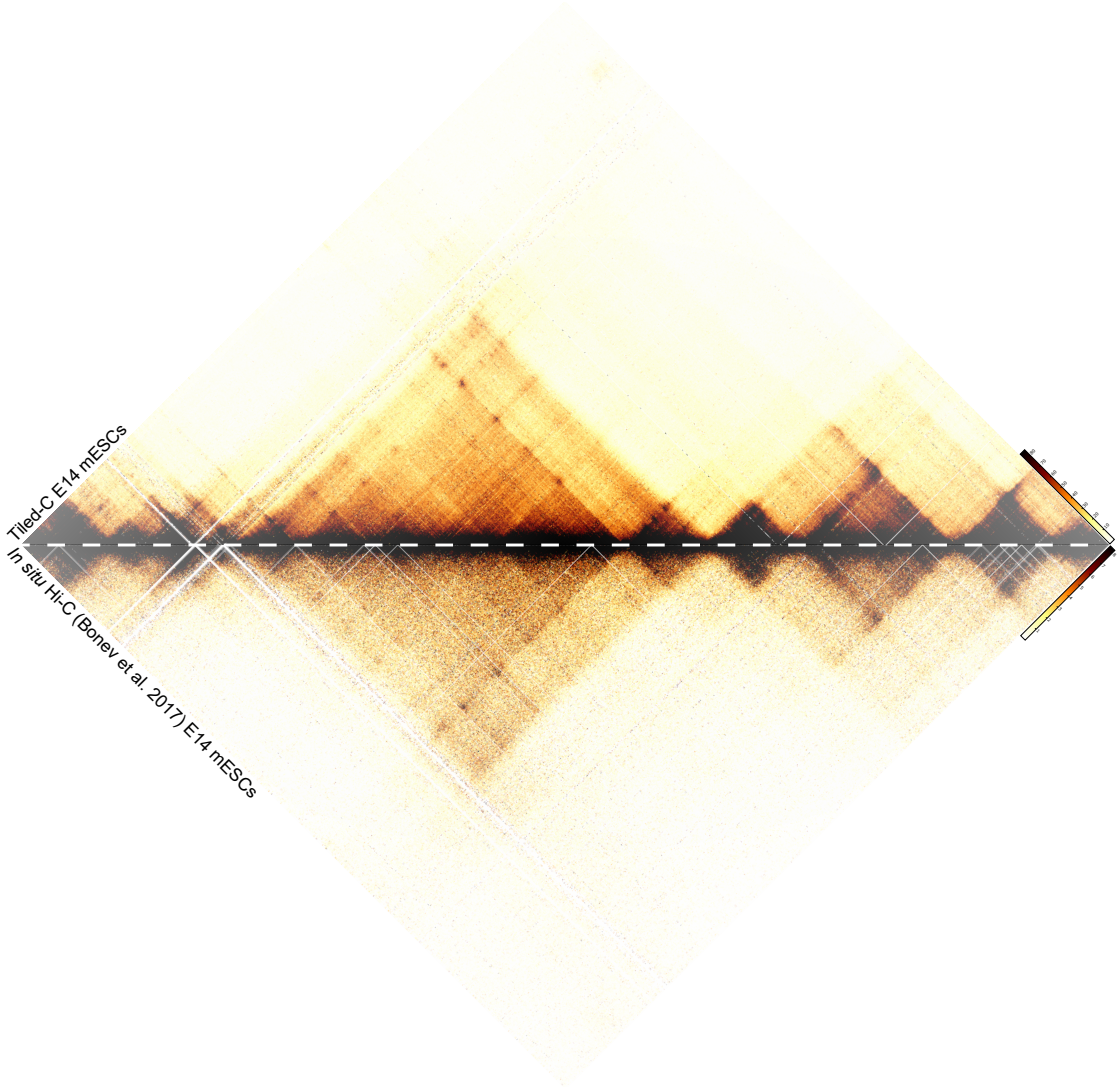


Figure 3.8 – Comparison between Tiled-C and *in situ* Hi-C Interaction matrices showing chromatin interactions in the *Runx1* domain in undifferentiated E14 mESCs are shown. The top matrix (above the white dashed line) depicts data generated by Tiled-C, while the bottom matrix depicts the highest resolution *in situ* Hi-C data to date (Bonev et al., 2017). Both data sets were analysed at 2 kb resolution and visualised with a threshold at the 94 th percentile. Hi-C data were analysed by Marieke Oudelaar.

3.2.5.2 Tiled-C reveals complex *cis*-interactions in 416B and E14 cells

Raw Tiled-C data was processed using a previously published normalisation method, Iterative Correction and Eigenvector decomposition (ICE, typically used for Hi-C data) (Imakaev et al., 2012), and scaled to adjust for sequencing depth. The resulting normalised Tiled-C *cis*-interaction matrices revealed a striking level of structure within the *Runx1* domain in both cell types. In mESCs, when the gene was minimally expressed, the overall 1.1 Mb interaction domain was visible as a TAD-like structure stretching from *Setd4/Cbr1* at the telomeric end of the gene desert to *Clic6/Rcan1* centromeric to *Runx1* (Figure 3.9 f and black outlined triangles in 3.10 f). Several smaller interaction domains (sub-TADs) with CTCF binding sites at their boundaries were visible within the larger domain (Figure 3.9 f and green dashed triangles in 3.10 f). The P2 promoter formed a sub-TAD extending to the promoters of *Clic6/Rcan1*, and both regions bound CTCF and Pol II in mESCs (Figure 3.9 e and arrowhead labelled '1' in 3.10 e). Discrete foci of interactions were also seen between the P2 promoter and multiple CTCF binding sites within the gene desert (Figure 3.9 e and green circles in 3.10 e). Thus, despite low *Runx1* expression in E14 mESCs, CTCF sites were engaging in long-distance *cis*-interactions.

Where *Runx1* was more highly expressed, in 416B cells, the overall TAD-like structure was less pronounced than in E14 mESCs, but still visible (Figure 3.9 a and black outlined triangle in 3.10 a). Compared to E14 mESCs, finer details within the TAD were more prominent in 416B cells. In particular, 'stripes' of interactions were seen emanating from the P2 promoter and -468 CTCF site (Figure 3.9 a-c and red dashed lines in 3.10 a-c). Along the 'stripes' of interactions, discrete foci of interaction could be seen between CTCF binding sites and both the *Runx1* promoters (Figure 3.9 a and green circles in 3.10 a-c). A 'cloud' of interactions was also seen at the top of the main TAD, reflecting long-range interactions between the two clusters of CTCF sites at the domain boundaries (Figure 3.9 a and large blue circle in 3.10 a). Another sub-TAD between the -468 CTCF site and *Setd4* was also prominent in 416B cells (Figure 3.9 a and arrowhead labelled '2' in 3.10 a).

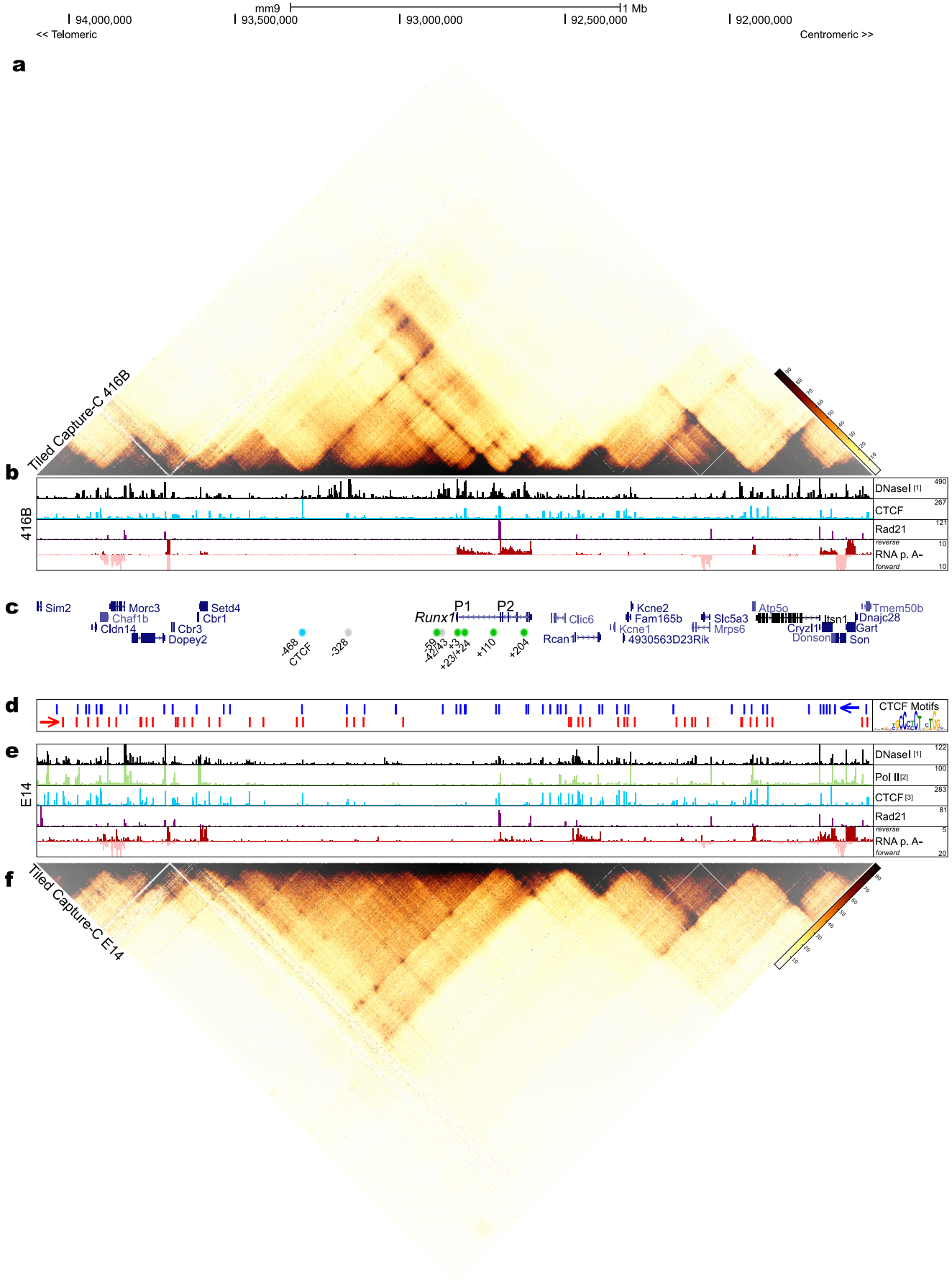


Figure 3.9 – Tiled-C analysis of 2.5 Mb around *Runx1* in 416B and E14 cells. *Legend continued on next page.*

Figure 3.9 – *Legend continued from previous page.* a and f) Normalised Tiled-C contact matrices at 2 kb resolution in 416B and E14 cells, respectively. b and e) Chromatin marks in 416B and E14 cells, respectively. Public data were 416B and E14 DNaseI ([1] Vierstra et al. 2014), E14 Pol II ([2], Rahl et al. 2010) E14 CTCF ([3] Handoko et al. 2011). c) Annotations showing coding genes (UCSC), the locations of *Runx1* enhancer elements previously identified (Nottingham et al., 2007; Bee et al., 2009b, 2010; Swiers et al., 2013a; Schütte et al., 2016) and shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles). The blue circle indicates the position of the -468 CTCF site in the gene desert. d) CTCF peaks called in 416B cells with *de novo* motif discovery. Blue and red bars indicate CTCF motifs oriented in the telomeric and centromeric direction, respectively. UCSC session of these data (Tiled data not supported by UCSC): `http://genome-euro.ucsc.edu/s/dowens/allData_DO_Tiled_main`

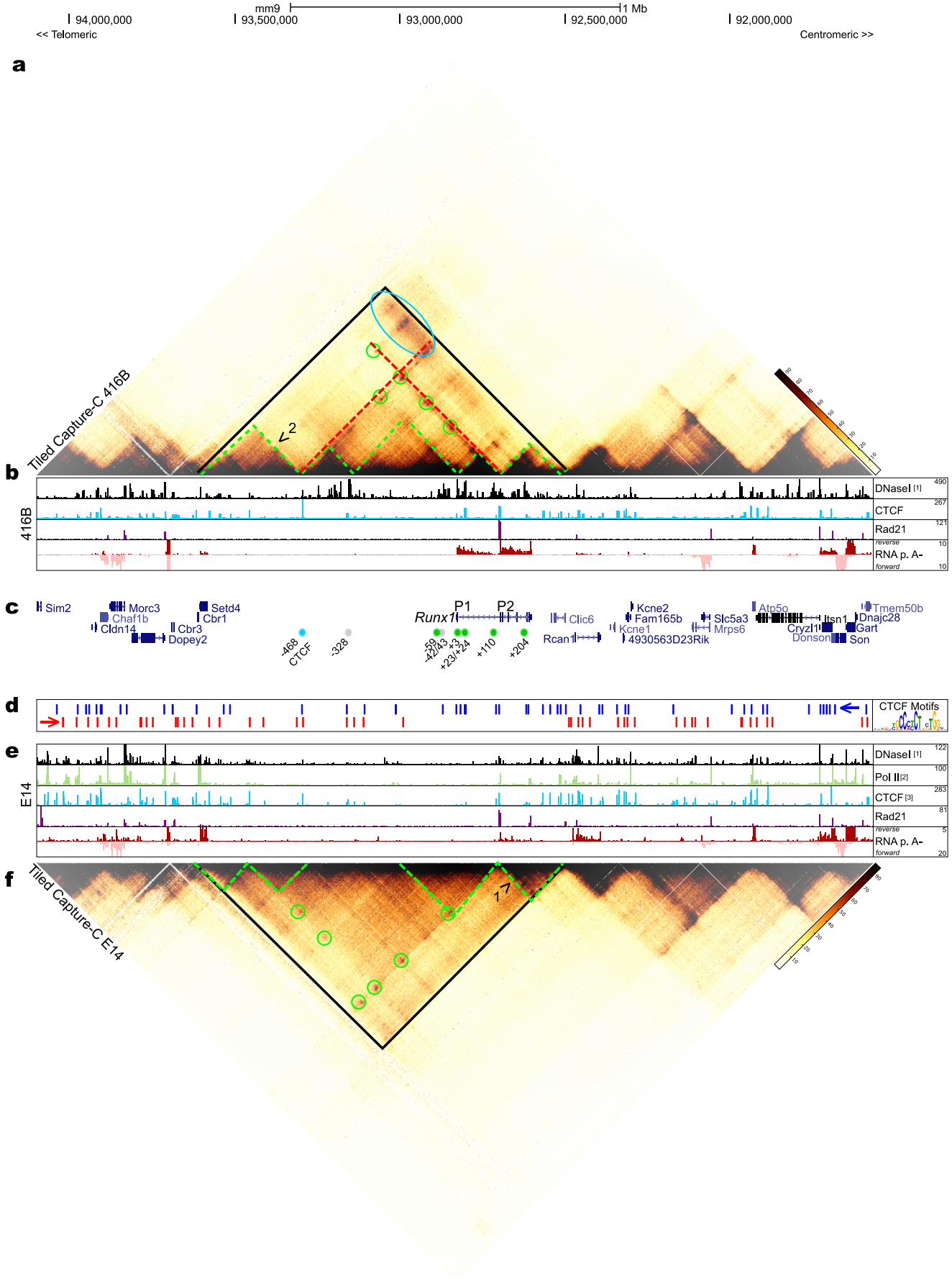


Figure 3.10 – Tiled-C analysis of 2.5 Mb around *Runx1* in 416B and E14 cells with additional labelling. *Legend continued on next page.*

Figure 3.10 – *Legend continued from previous page.* a and f) Normalised Tiled-C contact matrices at 2 kb resolution in 416B and E14 cells, respectively. b and e) Chromatin marks in 416B and E14 cells, respectively. The overall 1.1 Mb *Runx1* interaction domain (TAD) is highlighted by a black triangle. Smaller sub-TADs within the larger TAD are highlighted with dashed green triangles. 'Stripes' of interactions that were most prominent in 416B cells between the -468 CTCF site and *Runx1* P2 promoter are indicated with red dashed lines. Discrete foci of interactions that occurred between CTCF binding sites are indicated with green circles. Public data were 416B and E14 DNaseI ([1] Vierstra et al. 2014), E14 Pol II ([2], Rahl et al. 2010) E14 CTCF ([3] Handoko et al. 2011). c) Annotations showing coding genes (UCSC), the locations of *Runx1* enhancer elements previously identified (Nottingham et al., 2007; Bee et al., 2009b, 2010; Swiers et al., 2013a; Schütte et al., 2016) and shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles). The blue circle indicates the position of the -468 CTCF site in the gene desert. d) CTCF peaks called in 416B cells with *de novo* motif discovery. Blue and red bars indicate CTCF motifs oriented in the telomeric and centromeric direction, respectively. UCSC session of these data (Tiled data not supported by UCSC): `http://genome-euro.ucsc.edu/s/dowens/allData_DO_Tiled_main`

To visualise differences between the two cell types, I subtracted the matrices (Figure 3.11). After subtraction, the 'stripes' of interaction from the P2 promoter and -468 CTCF site were notably greater in 416B cells compared to E14 mESCs (Figure 3.11). The 'cloud' of interactions seen between the two boundary CTCF regions were increased in 416B cells compared to E14 mESCs (Figure 3.11). Moreover, discrete foci of interaction between CTCF sites were also increased in 416B cells compared to E14 (Figure 3.11). Statistical testing showed that interactions in the 'stripes', 'cloud', and foci were all significantly higher in 416B cells compared to E14 mESCs (Figure 3.12, a, FDR < 0.1). Compared to 416B cells, E14 cells showed significantly increased interaction frequencies diffusely throughout the overall TAD (Figure 3.12, f, FDR $<$ 0.1).

Upon close inspection of the Tiled-C data in the region of the *Runx1* gene itself, it was clear that each of the P1 and P2 promoters seemed to reside within two distinct sub-TADs (subcompartments) of frequent interactions (Figure 3.13 a and f). In 416B cells, both the P1 and P2 promoters appeared to be acting as loop boundaries at the corners of the two distinct sub-TADs (Figure 3.13 a). The boundary at the P1 promoter was prominent in 416B cells and absent in E14 mESCs (Figure 3.13 a and f). Indeed, subtracting the matrices and statistical testing revealed that interactions **between** the sub-TADs (i.e. across the P1 promoter and across the P2 promoter) were both significantly increased in E14 mESCs compared to 416B cells (Figures 3.14 f and 3.15 f, FDR < 0.1). Conversely, interactions **within** the sub-TADs were significantly greater in 416B cells compared to E14 mESCs (Figures 3.14 a and 3.15 a, FDR < 0.1). RNA-seq and CTCF ChIP-seq analysis showed that the 'boundary strength' of the promoters was correlated to promoter transcriptional output and CTCF binding at the promoters in each cell type (Figure 3.13 a, b, and e). Pol II binding was only assessed in E14 cells, and was observed at the P2 promoter but not at P1 (Figure 3.13 f). A notable exception to the P2 promoter seeming to act as a sub-TAD boundary within the gene was the interaction between the P1 promoter and 3'UTR of *Runx1* that was significantly higher in 416B cells compared to E14 mESCs (Figures 3.14 a and 3.15 a).

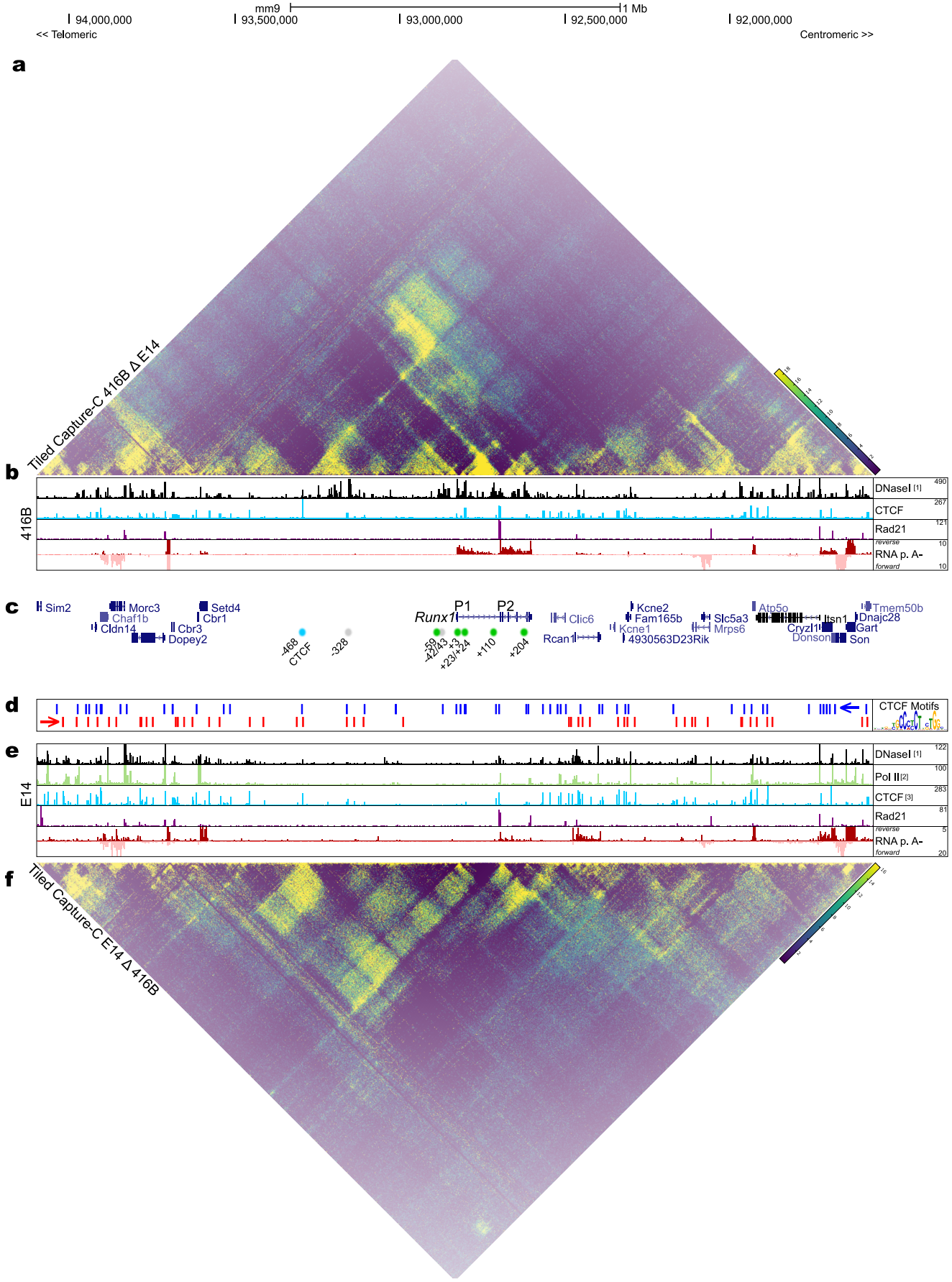


Figure 3.11 – Subtraction of Tiled-C matrices in 416B and E14 cells. *Legend continued on next page.*

Figure 3.11 – *Legend continued from previous page.* a) Normalised and subtracted (416B Δ E14) Tiled-C contact matrix at 2 kb resolution. b and e) Chromatin marks in 416B and E14 cells, respectively. Public data were 416B and E14 DNaseI ([1] Vierstra et al. 2014), E14 Pol II ([2], Rahl et al. 2010) E14 CTCF ([3] Handoko et al. 2011). c) Annotations showing coding genes (UCSC), the locations of *Runx1* enhancer elements previously identified (Nottingham et al., 2007; Bee et al., 2009b, 2010; Swiers et al., 2013a; Schütte et al., 2016) and shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles). The blue circle indicates the position of the -468 CTCF site in the gene desert. d) CTCF peaks called in 416B cells with *de novo* motif discovery. Blue and red bars indicate CTCF motifs oriented in the telomeric and centromeric direction, respectively. f) Normalised and subtracted (E14 Δ 416B) Tiled-C contact matrix at 2 kb resolution. UCSC session of these data (Tiled data not supported by UCSC): `http://genome-euro.ucsc.edu/s/dowens/allData_DO_Tiled_main`

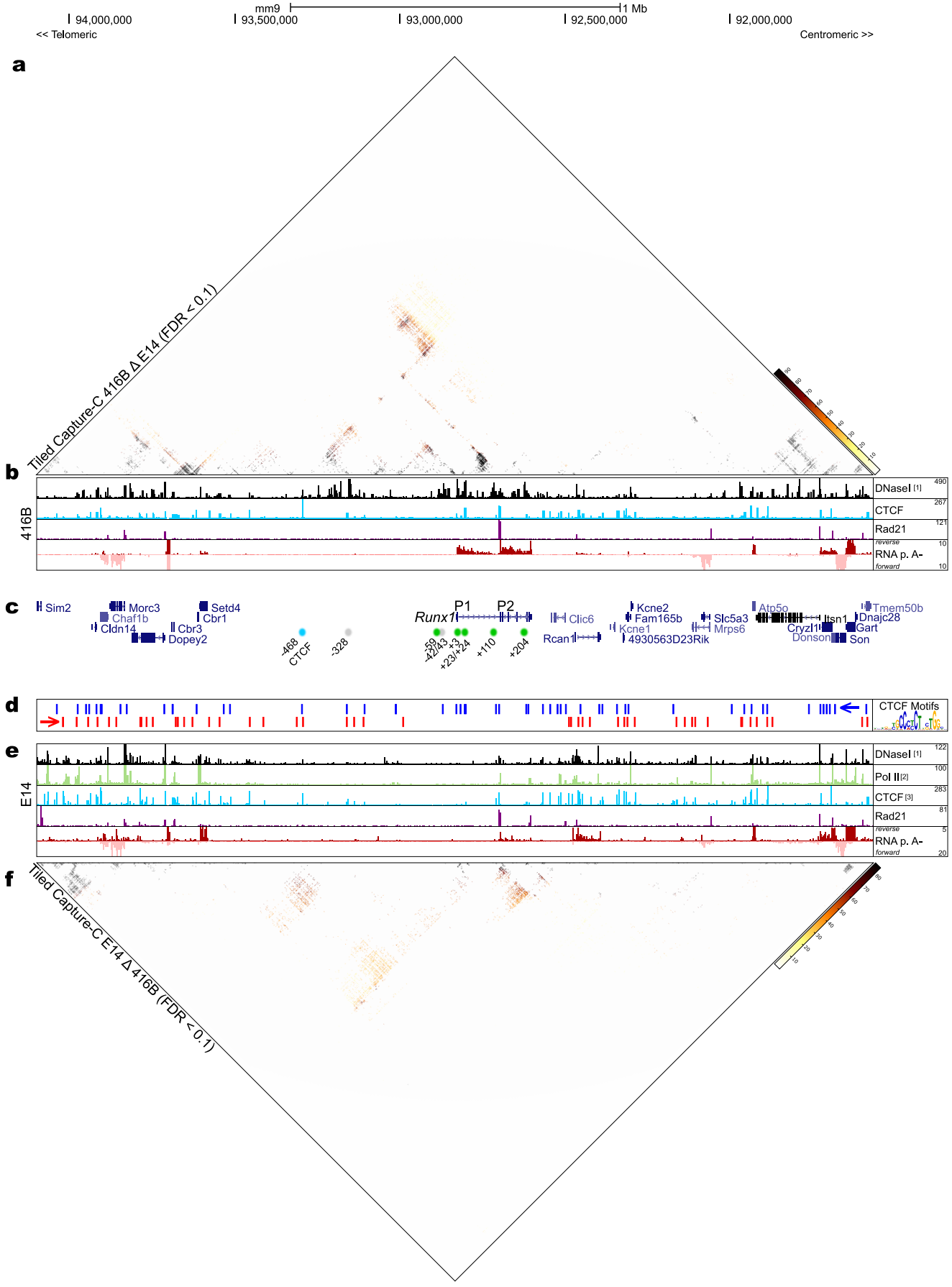


Figure 3.12 – Statistical testing of differences between Tiled-C matrices in 416B and E14 cells. *Legend continued on next page.*

Figure 3.12 – *Legend continued from previous page.* a and f) Normalised Tiled-C contact matrix at 2 kb resolution showing only the interactions that were significantly different between 416B and E14 cell types (FDR < 0.1). b and e) Chromatin marks in 416B and E14 cells, respectively. Public data were 416B and E14 DNaseI ([1] Vierstra et al. 2014), E14 Pol II ([2], Rahl et al. 2010) E14 CTCF ([3] Handoko et al. 2011). c) Annotations showing coding genes (UCSC), the locations of *Runx1* enhancer elements previously identified (Nottingham et al., 2007; Bee et al., 2009b, 2010; Swiers et al., 2013a; Schütte et al., 2016) and shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles). The blue circle indicates the position of the -468 CTCF site in the gene desert. d) CTCF peaks called in 416B cells with *de novo* motif discovery. Blue and red bars indicate CTCF motifs oriented in the telomeric and centromeric direction, respectively. UCSC session of these data (Tiled data not supported by UCSC): `http://genome-euro.ucsc.edu/s/dowens/allData_DO_Tiled_main`

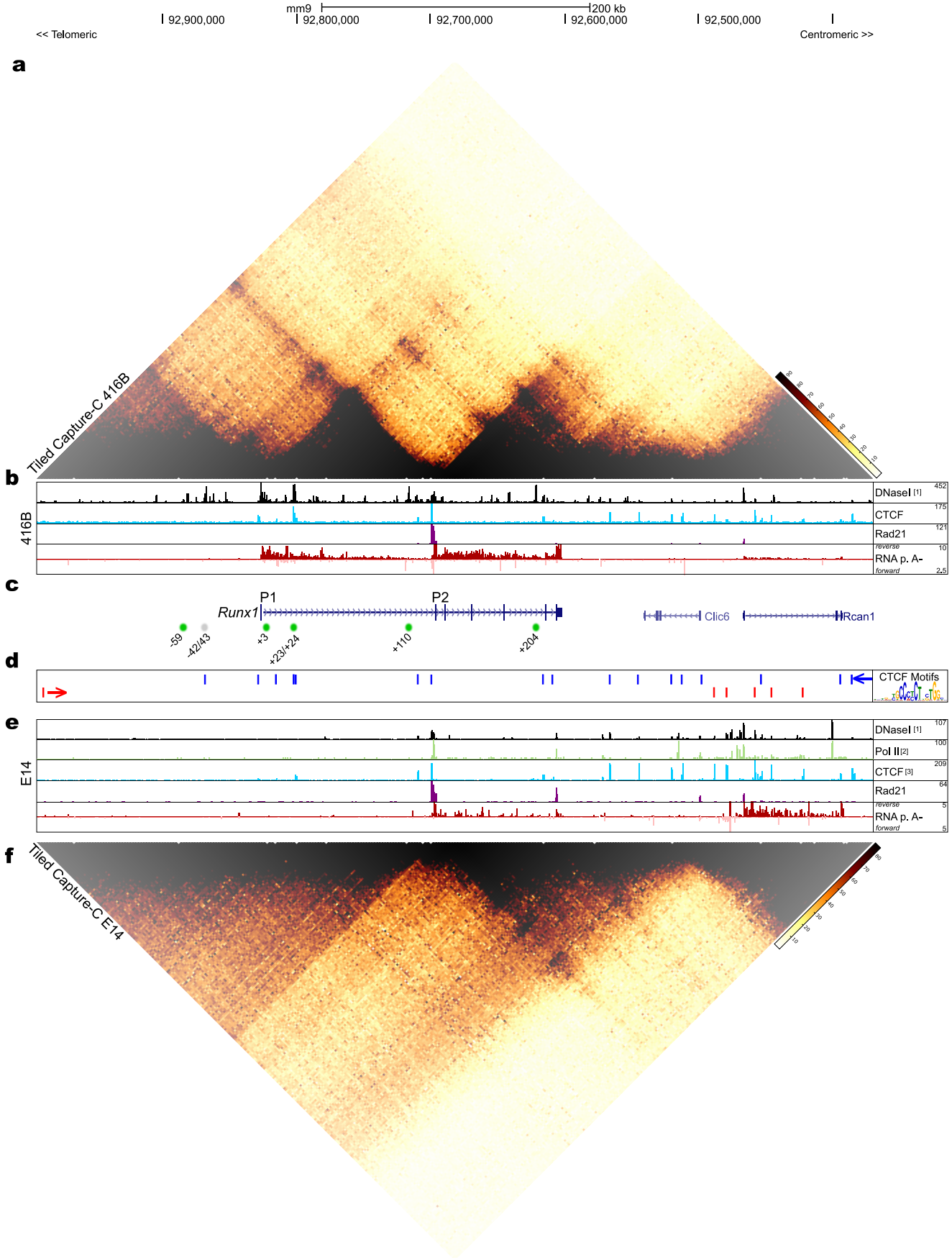


Figure 3.13 – Subcompartmentalisation within the *Runx1* gene in 416B and E14 cells.
Legend continued on next page.

Figure 3.13 – *Legend continued from previous page.* a and f) Normalised Tiled-C contact matrix at 2 kb resolution showing only the region surrounding the *Runx1*, *Clic6*, and *Rcan1* genes. b and e) Chromatin marks in 416B and E14 cells, respectively. Public data were 416B and E14 DNaseI ([1] Vierstra et al. 2014), E14 Pol II ([2], Rahl et al. 2010) E14 CTCF ([3] Handoko et al. 2011). c) Annotations showing coding genes (UCSC), the locations of *Runx1* enhancer elements previously identified (Nottingham et al., 2007; Bee et al., 2009b, 2010; Swiers et al., 2013a; Schütte et al., 2016) and shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles). d) CTCF peaks called in 416B cells with *de novo* motif discovery. Blue and red bars indicate CTCF motifs oriented in the telomeric and centromeric direction, respectively. UCSC session of these data (Tiled data not supported by UCSC): `http://genome-euro.ucsc.edu/s/dowens/allData_DO_Tiled_smallerWindow`

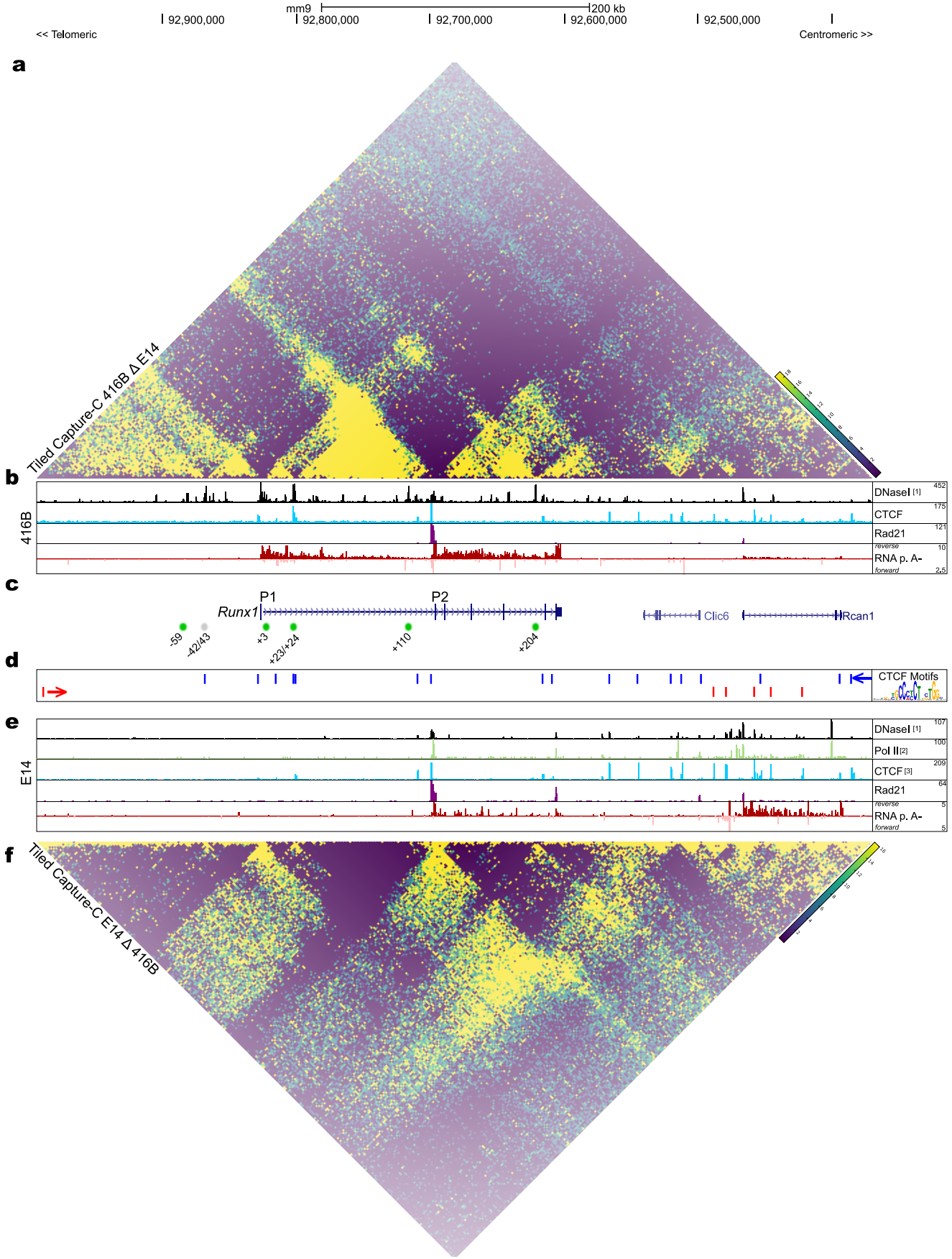


Figure 3.14 – Differences in subcompartmentalisation within the *Runx1* gene between 416B and E14 cells. *Legend continued on next page.*

Figure 3.14 – *Legend continued from previous page.* a) Normalised and subtracted (416B Δ E14) Tiled-C contact matrix at 2 kb resolution showing only the region surrounding the *Runx1*, *Clic6*, and *Rcan1* genes. b and e) Chromatin marks in 416B and E14 cells, respectively. Public data were 416B and E14 DNaseI ([1] Vierstra et al. 2014), E14 Pol II ([2], Rahl et al. 2010) E14 CTCF ([3] Handoko et al. 2011). c) Annotations showing coding genes (UCSC), the locations of *Runx1* enhancer elements previously identified (Nottingham et al., 2007; Bee et al., 2009b, 2010; Swiers et al., 2013a; Schütte et al., 2016) and shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles). d) CTCF peaks called in 416B cells with *de novo* motif discovery. Blue and red bars indicate CTCF motifs oriented in the telomeric and centromeric direction, respectively. f) Normalised and subtracted (E14 Δ 416B) Tiled-C contact matrix at 2 kb resolution. UCSC session of these data (Tiled data not supported by UCSC): `http://genome-euro.ucsc.edu/s/dowens/allData_DO_Tiled_smallerWindow`

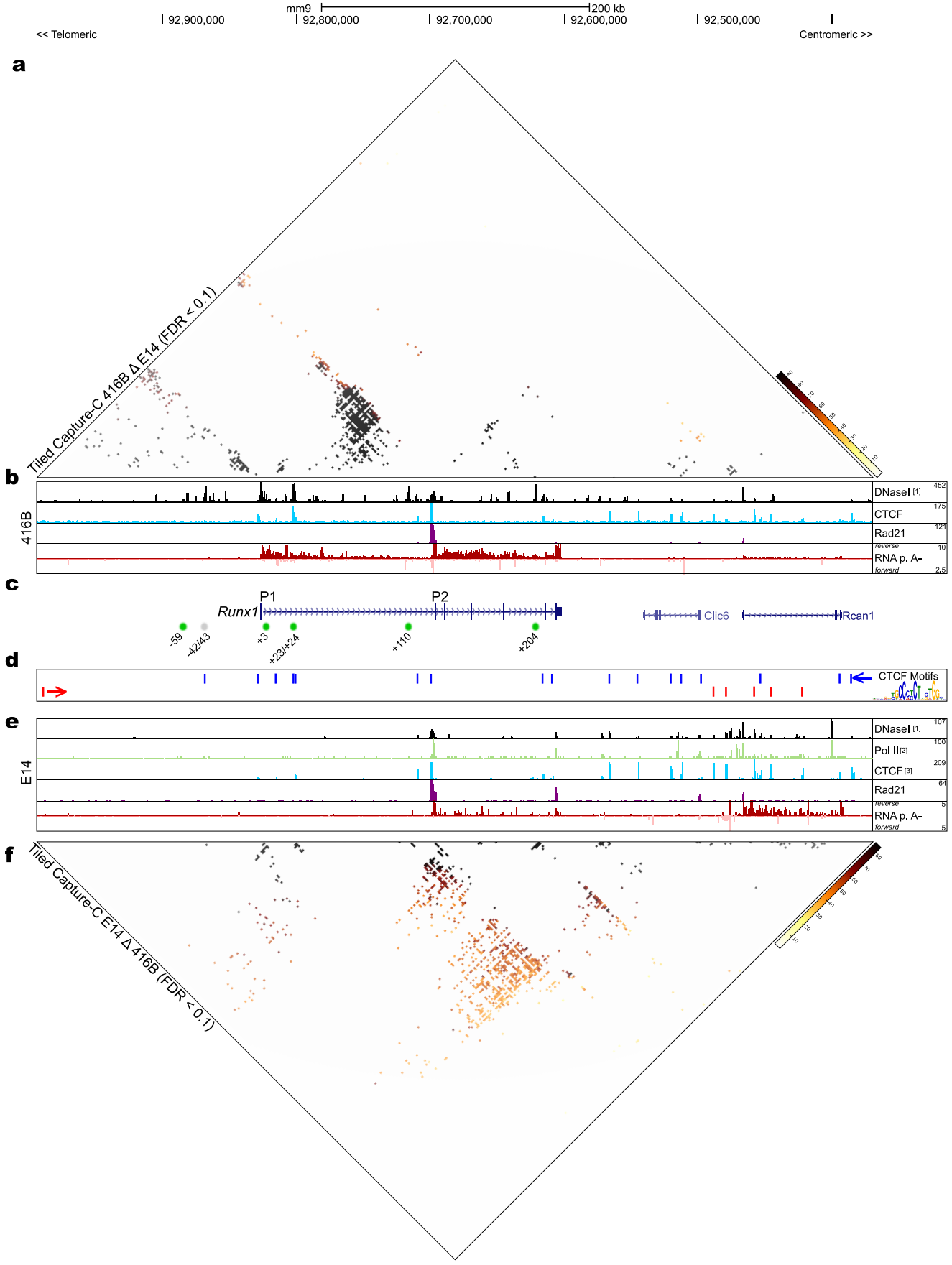


Figure 3.15 – Statistical testing of differences in subcompartmentalisation within the *Runx1* gene in 416B and E14 cells. *Legend continued on next page.*

Figure 3.15 – *Legend continued from previous page.* a and f) Normalised Tiled-C contact matrix at 2 kb resolution showing only the interactions that were significantly different between 416B and E14 cell types (FDR < 0.1) showing only the region surrounding the *Runx1*, *Clic6*, and *Rcan1* genes. b and e) Chromatin marks in 416B and E14 cells, respectively. Public data were 416B and E14 DNaseI ([1] Vierstra et al. 2014), E14 Pol II ([2], Rahl et al. 2010) E14 CTCF ([3] Handoko et al. 2011). c) Annotations showing coding genes (UCSC), the locations of *Runx1* enhancer elements previously identified (Nottingham et al., 2007; Bee et al., 2009b, 2010; Swiers et al., 2013a; Schütte et al., 2016) and shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles). d) CTCF peaks called in 416B cells with *de novo* motif discovery. Blue and red bars indicate CTCF motifs oriented in the telomeric and centromeric direction, respectively. UCSC session of these data (Tiled data not supported by UCSC): `http://genome-euro.ucsc.edu/s/dowens/allData_DO_Tiled_smallerWindow`

3.3 Chapter conclusions and discussion

The aim of this chapter was to explore differences in *cis*-interactions between cells where *Runx1* was transcriptionally inactive and active. Using Capture-C and Tiled-C I have defined *cis*-interactions in the *Runx1* domain at extremely high spatial resolution. This was done in a haematopoietic cell line that transcribes high levels of *Runx1* (416B cells) and undifferentiated mESCs (E14) that show low level *Runx1* expression. The overall *Runx1* promoter regulatory domain (TAD) was already visible in undifferentiated mESCs, implying that it forms independent of high levels of transcription. The TAD was $\sim$ 1.1Mb in size and spanned the entire adjacent gene desert. This defines the region within which the enhancers responsible for *Runx1* transcriptional regulation during EHT are most likely to be found, since enhancer-promoter interactions are thought to mainly only occur within TADs (Lettice et al., 2011; Symmons et al., 2014; Lupianez et al., 2015).

One prior study analysed *cis*-interactions in the *Runx1* domain using 4C-seq in another murine haematopoietic cell line, HPC-7 (Marsman et al., 2017). The *cis*-interacting domain of *Runx1* identified by Marsman et al. agrees well with the interaction domain I defined by Capture-C. However, by performing Capture-C in both E14 and 416B cells, I was additionally able to compare the changes in *cis*-interactions when the gene was transcribed at both high and low levels. This showed that *Runx1* enhancers interacted significantly more with both *Runx1* promoters when the gene was highly transcribed compared to when it was transcribed at low levels. This comparison between cell types has not been done before for *Runx1* and provides a strong indication that the *Runx1* enhancers previously identified (Nottingham et al., 2007; Bee et al., 2009b; Swiers et al., 2013a; Schütte et al., 2016) indeed regulate *Runx1* expression in haematopoietic cells. Analysis of chromatin marks in cells undergoing EHT is needed in order to address which specific enhancers are active during EHT. Moreover, genomic deletion of enhancers *in situ* needs to be performed in order to determine which of these enhancers may be functionally required. Both questions are addressed in chapter 4.

'Stripes' of chromatin interactions were seen, particularly in 416B cells when transcription levels of *Runx1* were high, that are consistent with cohesin complexes mediating pervasive interactions throughout a domain. Similar 'stripes' have previously been observed in Hi-C data and were associated with CTCF/cohesin binding sites close to active enhancers (Vian et al., 2018; Kraft et al., 2019; Barrington et al., 2019). A similar observation was made at *Runx1*–with Rad21 binding being highly enriched at the P2 promoter. It was surprising that Rad21 binding changed very little between 416B and E14 cells, since it has previously been observed that cohesin is associated with active tissue-specific enhancers (Kagey et al., 2010; Schmidt et al., 2010; Faure et al., 2012; Kojic et al., 2018). One possible explanation for the lack of Rad21 binding at active enhancers that I saw could be technical. It was previously observed that Rad21 binding at enhancers was considerably less enriched than Rad21 binding at CTCF sites (Faure et al., 2012). I had sufficient sequencing depth to identify strong Rad21 peaks at CTCF binding sites, and it is possible that with greater sequencing depth, less enriched Rad21 binding at enhancers would also become apparent. ChIP-seq for cohesin loading factor Nipbl could also be done, which is also associated with active enhancers (Kagey et al., 2010), and would allow cohesin

recruitment to be inferred more directly. Moreover, performing multiple replicates of ChIP-seq or ChIP-qPCR in each cell type would allow statistical comparison of Rad21 or Nipbl binding at the *Runx1* locus in the different cell types.

A 'cloud' of interactions was seen between the two broad boundary regions of the domain (that contain multiple convergently oriented CTCF binding sites) and these interactions were significantly increased in 416B cells compared with E14 mESCs. These diffuse interactions between clusters of boundary CTCF binding sites is consistent with loop extrusion bringing together CTCF sites at the base of a chromatin loop. Similar diffuse interactions around clusters of convergently oriented CTCF binding sites have been observed before at *α-globin* (Hanssen et al., 2017; Brown et al., 2018). By super-resolution microscopy, Brown et al. found that the internal part of the *α-globin* domain containing the *α-globin* gene promoters became 'decondensed' when transcription was active in erythroid cells (Brown et al., 2018), bringing the boundary CTCF sites into proximity. A similar 'decondensation' of the *Runx1* domain when transcription levels is consistent with my Capture-C and Tiled-C data, where the boundary CTCF sites interacted at a higher frequency in 416B cells compared to E14 mESCs. It is presently not clear what causes 'decondensation'. Transcription and enhancer-promoter interactions were not found to be necessary for the 'decondensed' *α-globin* domain to form (Brown et al., 2018), and one plausible candidate is increased cohesin-mediated loop extrusion.

Discrete subcompartments of preferential interactions were seen containing the two *Runx1* promoters, which correlated to transcriptional output from the promoters. This suggests that the active *Runx1* promoters might be able to act as barriers to loop extrusion. In pools of 416B cells, both promoters were active and seemed to form domain boundaries. Because all these assays were performed at the population level, however, it is possible that only one promoter is active at a time in any given cell, with only one sub-TAD formed. Single cell 3C assays like Tri-C (Oudelaar et al., 2018) and single cell RNA-seq could be used to address this. The fact that transcription from the *Runx1* promoters was associated with boundary formation agrees with several previous studies suggesting that actively transcribing promoters may act as boundaries (Hou et al., 2012; Dixon et al., 2012; Rao et al., 2014; Van Bortle et al., 2014; Moore et al., 2015; Bonev et al., 2017). Interestingly, CTCF is known to be associated with chromatin boundaries (Dixon et al., 2012; Rao et al., 2014; Dixon et al., 2015) and CTCF binding was also seen at the *Runx1* promoters, suggesting that CTCF may also be responsible for sub-TAD boundary formation at the *Runx1* promoters.

It was interesting that the sub-TAD boundary at P2 was visible in both E14 and 416B cells, while the P1 promoter sub-TAD boundary was only observed in 416B cells. Low levels of transcription, Pol II and CTCF binding were seen at P2 in E14 mESCs, which was surprising because pluripotent cells generally lack lineage specific TF expression (Handyside et al., 1989). Therefore, the observed sub-TAD boundary at P2 in E14 mESCs could be due to CTCF binding or low levels of transcription. The low levels of transcription from P2 observed in mESCs could be due to P2 being a poised and bivalently marked polycomb repressed promoter, which are typically associated with developmentally regulated TF genes (Bernstein et al., 2006), like *Runx1*.

An alternative explanation for this low level of expression could be spontaneous differentiation which could lead to sporadic *Runx1* expression in small numbers of mESCs. Flow cytometry of the Runx1-Venus mESC line or single cell RNA-seq analysis could be used to distinguish these two scenarios. The relationship between CTCF binding and promoter activity will be further explored in chapter 5.

Capture-C combined with CTCF ChIP-seq showed that multiple CTCF binding sites form a gradual boundary that delimits the extent of the *Runx1* domain. This was apparent because *cis*-interactions decreased in a stepwise manner when looking from viewpoints within the *Runx1* domain (i.e. from promoters or enhancers). *De novo* motif discovery showed that the two clusters of CTCF sites at either end of the *Runx1* interaction domain are convergently oriented towards each other. This arrangement of convergently oriented CTCF motifs has previously been suggested to play a role in boundary formation, plausibly via CTCF/cohesin-mediated loop extrusion (Rao et al., 2014; Vietri Rudan et al., 2015; Guo et al., 2015; de Wit et al., 2015). It seems likely that multiple CTCF sites together at the *Runx1* domain boundaries are acting redundantly to form the boundary. Indeed, previous work showed that deletion of two adjacent CTCF sites was required to expand the interaction domain containing the *α-globin* enhancers and promoters (Hanssen et al., 2017; Oudelaar et al., 2018). Multiple successive CTCF binding sites may be required to form a robust chromatin boundary in order to account for the fact that a single CTCF site may not be bound by CTCF at a given moment in a given cell. Indeed, CTCF was shown to have a short residency time on chromatin of only one or two minutes (Hansen et al., 2017). The presence of multiple CTCF binding sites within a boundary could also potentially buffer against deleterious mutations during evolution leading to loss or inversion of a single CTCF binding site.

Neighbouring genes to *Runx1* generally resided outside of the cluster of CTCF sites that make up the boundaries of the *Runx1* domain despite being contained within a shared conserved block of synteny (Ahituv et al., 2005). However, *Runx1* and *Clic6* did interact with each other significantly in mESCs, suggesting that they might be involved in regulating each other. Interestingly, two of the centromeric boundary CTCF sites of the *Runx1* interaction domain are located within the introns of *Clic6*. Therefore, the observed interaction between *Runx1* and *Clic6* could be in part mediated by these CTCF sites. Alternatively, both sites were bound by Pol II, which has previously been associated with chromatin loops present during development (Ghavi-Helm et al., 2014). Deletion of the boundary CTCF sites within the *Clic6* gene could be done to examine the role of these CTCF sites in the observed *Runx1-Clic6* interaction in mESCs. It would also be interesting to see whether they play a role in regulating the expression of *Runx1* or *Clic6*. The fact that most of the genes with the syntenic block did not interact frequently with *Runx1* suggests that they are unlikely to share *cis*-regulatory elements. Why, then, would a block of synteny be evolutionarily conserved? One possibility is that in another cell type not examined here the genes do share *cis*-regulatory mechanisms.

The *Runx1* domain boundaries bordered by convergent CTCF sites are patently not absolute, as interactions were observed between *Runx1* enhancers and *Erg*, 3 Mb away. Marsman et al. also observed similar interactions between +23/+24 enhancer and

Erg by 4C-seq (Marsman et al., 2017). The authors suggested that the +23 enhancer '*was highly interactive... with loci outside the* [Runx1] *TAD*' and might be taking part in '*the formation of a local active chromatin hub controlling* Runx1 *expression in haematopoietic cells*' (Marsman et al., 2017). Capture-C provides an accurate quantification of the frequency of these interactions (by removing PCR duplicates and only counting **unique** *cis*-interactions). This analysis showed that interactions between the *Runx1* enhancers and *Erg* (and interactions between the *Runx1* promoters and *Erg*) were two orders of magnitude less frequent than between the *Runx1* enhancers and *Runx1* promoters. This suggests that these long-range interactions are unlikely to play a major role in transcriptional regulation of the genes, but this remains to be determined experimentally. It is also unclear how the *Runx1* enhancers and *Erg* might be locating each other over 3 Mb genomic linear distance. Such long-distance interactions are unlikely to be maintained by CTCF/cohesin-mediated loop extrusion, due to the presence of fifty or so CTCF binding sites between the *Runx1* and *Erg* loci. It seems likely, therefore, that these interactions reflect a higher-order condensate of multiple transcriptionally active elements possibly maintained by a mechanism such as phase separation (Hnisz et al., 2017). This notion is supported by the fact that both genes were highly expressed, and contained regions ranked as some of the top 'super-enhancers' in 416B cells.

4. Dynamic enhancer activation and functional requirements during endothelial-to-haematopoietic transition

4.1 Introduction

During development, the expression of *Runx1* is subject to tight spatiotemporal transcriptional control. Multiple haematopoietic *Runx1* enhancers have been identified by our group and others and their activity characterized in enhancer-reporter assays *in vitro* and *in vivo* (Nottingham et al., 2007; Bee et al., 2009b, 2010; Ng et al., 2010; Swiers et al., 2013a; Schütte et al., 2016; Marsman et al., 2017) (Figure 1.4 a). The *Runx1* +23 and +204 enhancer-reporters are active first during EHT; +23 enhancer-reporter activity is seen in HE cells in the yolk-sac blood islands from E7.5 (Nottingham et al., 2007; Bee et al., 2010; Swiers et al., 2013a), and +204 enhancer-reporter is seen in HE cells from E8.5 (de Bruijn lab, unpublished observations). Further down the EHT trajectory, the +110 enhancer-reporter becomes expressed in HP cells (de Bruijn lab, unpublished observations). This shows that the *Runx1* enhancers in isolation exhibit distinct and partially overlapping spatiotemporal activities during EHT. However, it remains unclear whether the activities of the haematopoietic *Runx1* enhancers in enhancer-reporter constructs is identical to those of the enhancers in their endogenous locus. It is also not known what upstream factors might mediate the cell type-specific activities of the *Runx1* enhancers.

It has been suggested that multiple functionally redundant 'shadow enhancers' are a common feature in the transcriptional regulation of developmentally regulated genes (Cannavò et al., 2016; Osterwalder et al., 2018), but it is unclear to what extent the *Runx1* haematopoietic enhancers previously identified (Nottingham et al., 2007; Schütte et al., 2016) may be functionally redundant. Moreover, *Runx1* haematopoietic enhancers have previously been suggested to act synergistically as a SE (Mill et al., 2019; Gunnell et al., 2016; Schuijers et al., 2018; Hnisz et al., 2017; Saint-André et al., 2016; Kwiatkowski et al., 2014), but this possibility has so far not been examined during EHT.

In this chapter, endogenous *Runx1* enhancer activation is investigated by assaying chromatin accessibility during EHT *in vitro*. Candidate upstream regulators of *Runx1* enhancers during EHT are identified by examining TF binding and performing digital DNaseI footprinting (see section 1.4.3) in cellular intermediates of EHT. Finally, possible redundancy and synergy between *Runx1* enhancers is examined by deleting enhancers in the endogenous *Runx1* locus using CRISPR/Cas9 in mESCs. After developing a robust genotyping strategy to detect unwanted larger deletions (LDs), functional enhancer requirements are examined by differentiating enhancer knock-out

lines *in vitro* to generate cellular intermediates of EHT.

4.2 Results

4.2.1 Recapitulating developmental haematopoiesis *in vitro*

To model developmental haematopoiesis in vitro I made use of a +23Cherry::Runx1Venus double reporter mESC line (23C-RV) that had just been generated by Lucas Greder, who at the time also was a DPhil student in the lab. This line contained mCherry under the control of the +23 enhancer in the *Col1a1* safe harbour locus and a Venus reporter at the 3' end of Runx1 resulting in a functional Runx1-P2A-Venus protein (Figure 4.2 a). Using a chemically defined *in vitro* differentiation protocol (Figure 4.1), Lucas and I differentiated 23C-RV cells to test their behaviour and reporter expression in haematopoietic development *in vitro*. We isolated distinct cellular populations undergoing EHT *in vitro* (Figure 4.2 b). After 4 days of embryoid body (EB) culture (Figure 4.2 b and c), 23C-RV cells robustly generated Flk1+ mesodermal progenitors (Figure 4.2 b and f) that were FACS purified. After replating in a haemogenic cytokine cocktail, significant numbers of Flk1+ cells underwent EHT and differentiated towards haematopoietic lineages. Colonies of CD31 endothelial cells were visible (outlined with yellow dashed lines in Figure 4.2 b and c) that contained Runx1- endothelial cells (marked with an orange star in Figure 4.2 d) as well as Runx1+ HE cells (marked with a solid white arrowhead in Figure 4.2 c). Rounded HP cells that were CD31+ CD41+ Runx1+ could be seen budding off from the HE layer, indicating that successful EHT took place within the cultures (hollow white arrow heads in Figure 4.2 c and d). Colony formation assays revealed that functional haematopoietic progenitors were produced (Figure 4.2 e). The 23Cherry enhancer-reporter allele was weak in imaging experiments but flow cytometry revealed that Ter119-, VE-Cadherin+, CD41-, CD45- 23Cherry+ HE cells were generated in the cultures (Figure 4.2 f). FACS controls are shown in Appendix Figures 7.2 and 7.3. Thus the *in vitro* EHT differentiation of the 23C-RV cells allowed us to isolate sequential stages of developing HE, that is Runx1- 23Cherry+ competent HE cells (cHE), and Runx1+ specifying HE cells (sHE), as well as and Ter119- VE-Cadherin+ CD41+ haematopoietic progenitor (HP) cells that were almost exclusively 23Cherry+ and *Runx1*-Venus+ (Figure 4.2 f).

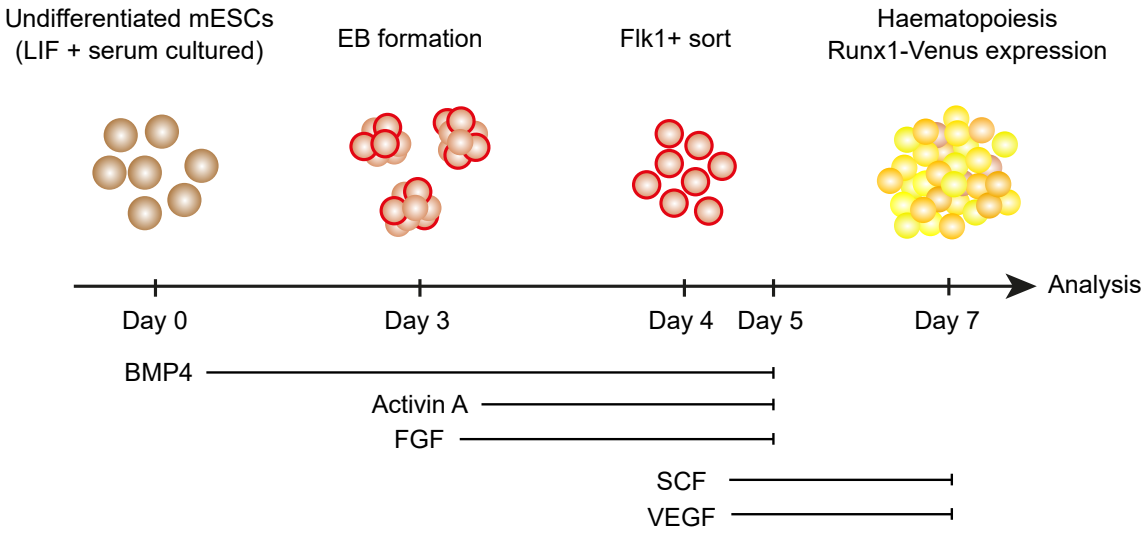


Figure 4.1 – Schematic of the *in vitro* haematopoietic differentiation protocol adapted by Lucas Greder and myself. After (Sroczynska et al., 2009b; Pearson et al., 2015). Cytokines used are labelled underneath the days when they were added to the chemically defined medium.

4.2.2 *Runx1* enhancers are dynamically activated during *in vitro* EHT

We performed ATAC-seq using a low cell number-adapted protocol (Maria Suciu, Hughes lab, MRC MHU) to assess chromatin accessibility in populations of 50,000 purified cHE, sHE, and HP cells. I performed an initial analysis and inspection of the data which revealed a low signal-to-noise ratio particularly in the cHE and sHE samples (data not shown). To circumvent this problem and maximise signal and reduce noise, windowing, normalisation, and merging of the cHE and sHE samples into a single HE sample (containing 23Cherry+ Runx1-Venus+/- HE) was performed. After these additional data processing steps, 119,340,022 uniquely mapping reads for HE cells (4 biological samples, 7 technical replicates) and 28,104,222 uniquely mapping reads for HP cells (1 biological sample, 2 technical replicates) were retained. To corroborate our data, I also analysed a publicly available DNaseI-seq data set from mESC-derived HE (CD41- Tie2+ Kit+) and HP (CD41+) cells generated previously (Goode et al., 2016). Quality control of both data sets revealed signal enrichment at the transcription start site (TSS) of expressed genes including *Polr2a* (Figure 4.3 a), indicating acceptable data quality. The promoters of pluripotency factors like *Pou5f1* were only accessible in undifferentiated E14 mESCs (Figure 4.3 b), indicating that the majority of the cells differentiated and lost expression of pluripotency-associated genes. Meta-plotting enrichment at all TSSs showed significant enrichment that was comparable between our ATAC-seq data and the publicly available DNase-seq data sets (Goode et al., 2016). Together, this suggested that these data sets would be useful for interrogating *Runx1* enhancer activities during EHT *in vitro*.

Analysis of our ATAC-seq data over the *Runx1* locus showed several peaks changing in accessibility in cells undergoing *in vitro* EHT (Figure 4.4 a). Enrichment of ATAC-seq reads at +23 was seen in 23Cherry+ HE cells and enrichment increased in CD41+ HP cells, indicating an increase in activity of the enhancer (Figure 4.4 a). Enrichment at

+110, however, was only seen in HP cells. +204 was slightly accessible in 23Cherry+ HE cells, and this was similar or subtly decreased in CD41+ HP cells (Figure 4.4 a). The same enhancer regions also interacted significantly with both *Runx1* promoters by Capture-C in 416B cells, further supporting their role as *Runx1* enhancers active during EHT (FDR < 0.1, Figure 4.4 b). DNaseI-seq in HE (CD41- Tie2+ Kit+) and HP (CD41+) cells generated previously (Goode et al., 2016) showed a similar profile of chromatin accessibility changes from HE to HP cells to our ATAC-seq data (Figure 4.4 a). Publicly available RNA-seq generated in the same cells as the DNaseI-seq data (Goode et al., 2016) confirmed the expected increase in transcription of *Runx1* from HE to HP cells (Appendix Figure 7.4). In agreement with the chromatin accessibility and expression changes, publicly available ChIP-seq showed H3K27ac enrichment was increased at +23 and +110 enhancers from HE to HP cells, and decreased at +204. Together, this shows that dynamic chromatin accessibility changes occur at *Runx1* enhancers during EHT *in vitro*. Together, this shows that *Runx1* enhancers are dynamically active in the endogenous *Runx1* locus correlated to an increase in its expression during EHT *in vitro*.

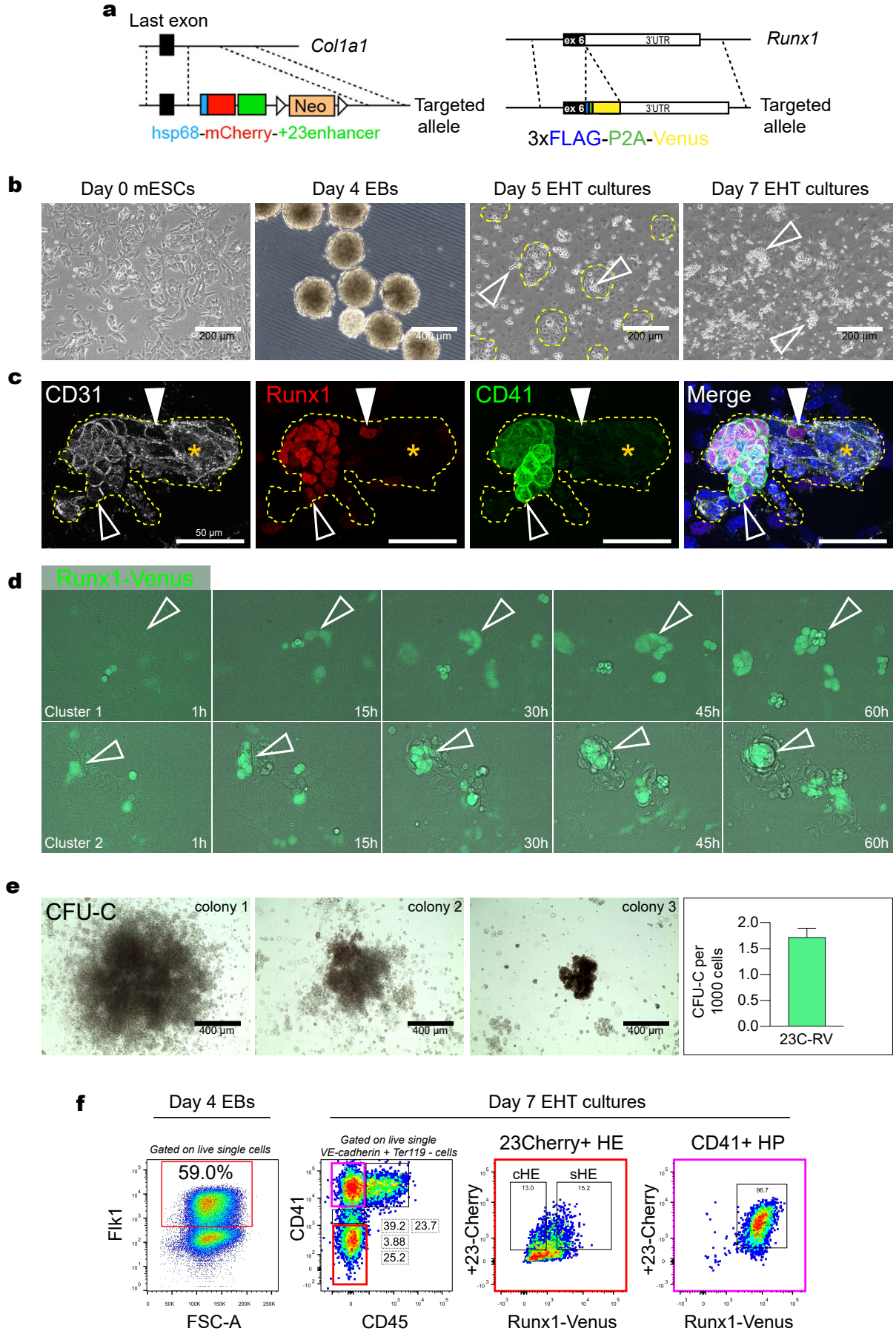


Figure 4.2 – *In vitro* differentiation of mESCs to haematopoietic lineages. *Legend continued on next page.*

Figure 4.2 – *Legend continued from previous page.* a) Schematic of targeting strategy designed by Lucas Greder to generate 23C-RV cells. +23Cherry was knocked into the *Col1a1* safe harbour locus and Venus was knocked in before the 3'UTR of both *Runx1* loci (homozygous). b) Phase-contrast images showing cells at the different stages of *in vitro* haematopoietic differentiation indicated. c) Imunnofluorescence imaging showing a colony of cells (outlined by a yellow dashed line) containing CD31+ CD41- Runx1- endothelial cells (orange star), CD31+ CD41- haemogenic endothelial cells (HE, solid white arrowhead), and CD31+ CD41+ Runx1+ haematopoietic progenitor cells (HP, white outlined arrowhead). d) Live imaging of 23C-RV cells performed by Joe Harman over 60 hours showing Runx1-Venus expression being upregulated in cells undergoing EHT. Outlined white arrow heads indicate the location of one cluster of Runx1-expressing cells. e) Colony forming potential from day seven haematopoietic differentiation cultures (n=3). f) Reanalysis and representative FACS plots from one differentiation showing percentage of Flk1+ mesodermal cells in disaggregated embryoid bodies at day 4. At day 7, percentages of CD41-, CD41 low, CD41+, and CD45 populations are shown. CD41- cells were further gated and sorted on 23-Cherry+ *Runx1*-Venus- (cHE) and 23-Cherry+ *Runx1*-Venus+ (sHE) and percentages of these populations are shown. Almost all sorted CD41+ HP cells were also 23-Cherry+ *Runx1*-Venus+.

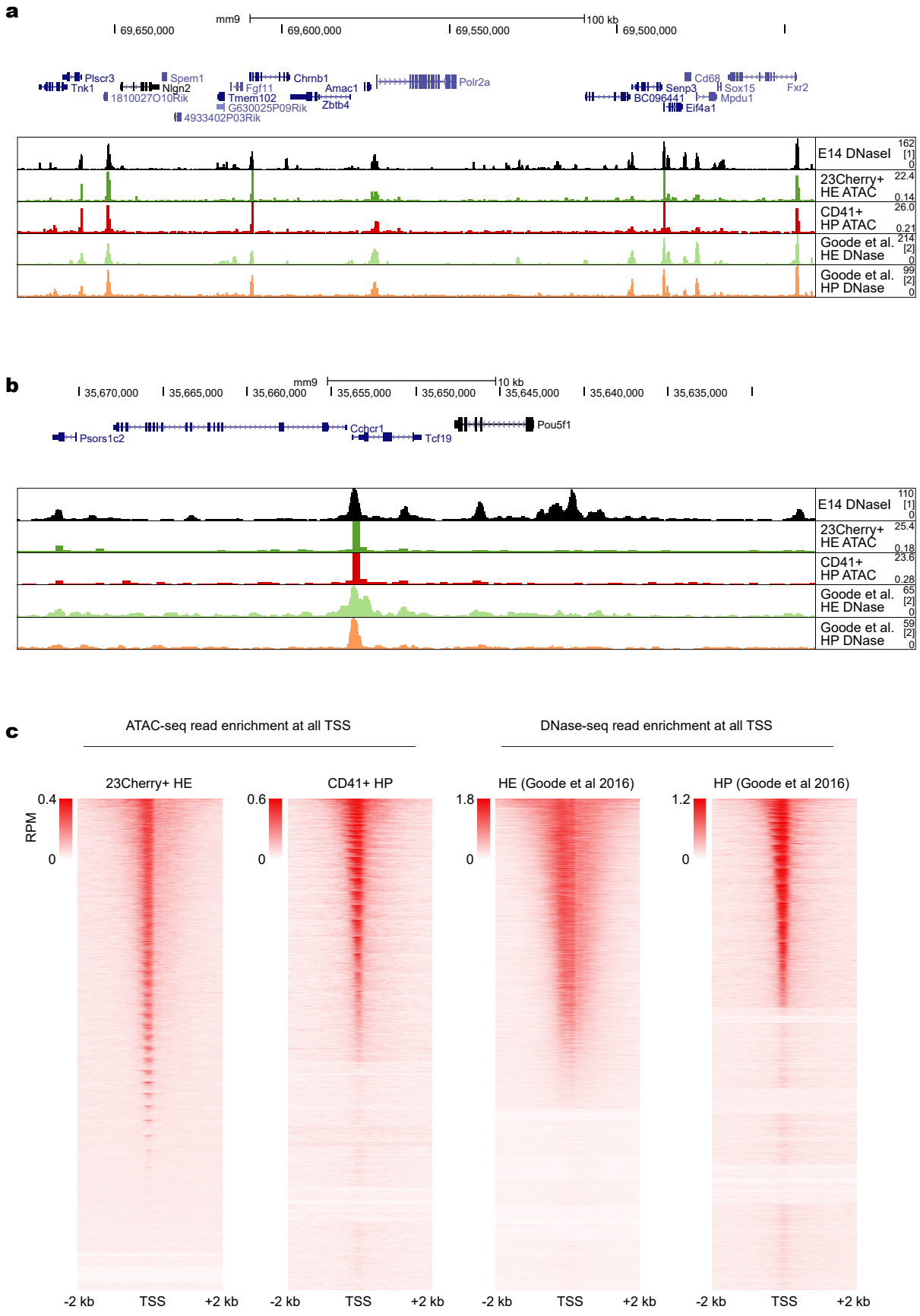


Figure 4.3 – Quality control of ATAC-seq data generated in populations undergoing EHT *in vitro Legend continued on next page.*

Figure 4.3 – *Legend continued from previous page.* ATAC-seq data generated in 23Cherry+ HE and CD41+ HP cells by Lucas Greder and myself and analysed by me. Public data analysed by me were DNaseI-seq generated in undifferentiated E14 mESCs ([1] Vierstra et al. 2014), and DNase-seq generated in HE and HP ([2] Goode et al. 2016). a) Reads were enriched in all samples at the promoters of expressed genes like *Polr2a*. b) Peaks were not seen at pluripotency associated genes like *Pou5f1*. c) Heatmap showing pile-up of reads across 4kb centred on all TSSs. UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_ATAC_mESC_Goode_QC`.

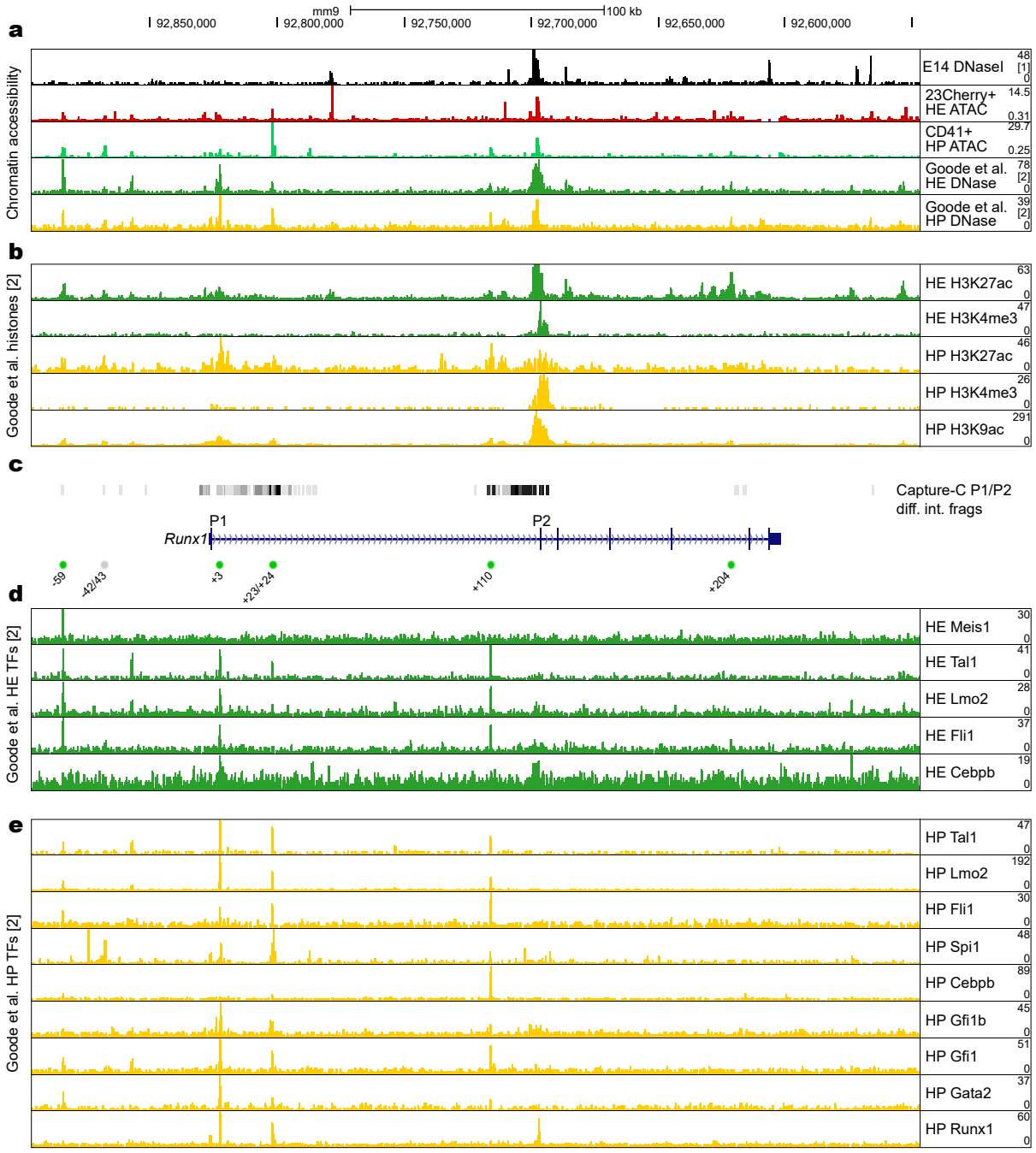


Figure 4.4 – Dynamic activation of *Runx1* enhancers in populations udergoing EHT *in vitro*. a) Chromatin accessibility in undifferentiated mESCs and mESC derived HE and HP cells. DNaseI-seq in undifferentiated E14 mESCs ([1] Vierstra et al. 2014), ATAC-seq in 23Cherry+ HE and CD41+ HP, DNase-seq generated in HE and HP ([2] Goode et al. 2016). b) Histone marks in mESC-derived HE and HP cells ([2] Goode et al. 2016). c) Annotation showing *Runx1* gene, P1 and P2 promoters, enhancers previously shown to drive reporter expression in haematopoietic (green circles) and non haematopoietic active sites (grey circles) in embryos. Regions that interacted significantly more with *Runx1* promoters in 416B cells compared to undifferentiated E14 mESCs by Capture-C are shown (diff. int. frag). d and e) TF binding in mESC-derived HE and HP cells ([2] Goode et al. 2016). UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_ATAC_mESC_Goode`

4.2.3 Upstream factors binding to *Runx1* enhancers during endothelial-to-haematopoietic transition

To assess TF binding at *Runx1* enhancers during *in vitro* EHT, I analysed publicly available ChIP-seq data for common haematopoietic TFs in HE (CD41- Tie2+ Kit+) and HP (CD41+) cells (Goode et al., 2016). TF binding was prominent at -59, +3, +23, and +110 enhancers which correlated with H3K27ac enrichment (Figure 4.4 b and d). Upstream factors binding these enhancers in common in both cell types included Scl/Tal1, Lmo2, and Fli1 (Figure 4.4 d). Differential TF binding was seen for Meis1, however, which only bound at -59 in HE cells (ChIP not done in HP cells, Goode et al. 2016). In both HE and HP cells, Runx1 bound at +3, +23, and the P2 promoter but was absent from the +110 enhancer, which is known to lack any Runx1 motifs (Schütte et al., 2016). Cebpβ was not strongly enriched at any enhancer in HE cells and bound only to the +110 enhancer in HP cells (Figure 4.4 d). Overall, multiple TFs bound to Runx1 enhancers during EHT that was correlated with their activity. Interestingly, despite showing limited marks of transcriptional activity in HE cells, +110 enhancer showed significant peaks for TFs Scl/Tal1, Lmo2, and Fli1 binding (Figure 4.4 d), possibly suggesting these factors were binding before the enhancer was fully active. The +204 enhancer on the other hand, showed H3K27ac enrichment in HE cells but none of the TFs assayed here bound (Figure 4.4 d). This suggested that other hitherto unexamined TFs may also be contributing to *Runx1* enhancer activity.

4.2.4 Identification of upstream factors regulating *Runx1* enhancers during endothelial-to-haematopoietic transition

To investigate novel upstream TFBS that might be responsible for the observed cell-type-specific differences in *Runx1* enhancer activities, I used a combination of bioinformatic approaches. ATAC-seq is currently not compatible with footprinting due to significant cutting biases (Schwessinger et al., 2017). Instead, publicly available DNaseI-seq data from HE (Goode et al., 2016) cells were used for footprinting. The quality of DNaseI-seq data generated in HP cells was below the threshold required to generate robust footprints, and so data from 416B cells (Vierstra et al., 2014) were used as a substitute. Using Sasquatch, meta DNaseI footprints were plotted across each enhancer in 7 bp windows (Schwessinger et al., 2017). Each 7-mer was then mutated *in silico* in a base-wise manner and DNaseI footprints were recalculated. By comparing the original and *in silico* mutated footprints, sequences that contained a significant footprint were identified ('damage score hits', Schwessinger et al. 2017). A damage score hit suggests a DNaseI footprint was sensitive to *in silico* mutation and are enriched for sequences with regulatory potential (Schwessinger et al., 2017). Many damage score hits were seen across +23, +110, and +204 enhancers (damage score > 1.0, Figure 4.5). Many of the hits identified were similar between HE and 416B cells, while others were unique to each cell type (compare green and red boxes in Figure 4.5). Overlaying consensus TFBS motifs (Heinz et al., 2010) and multi-species sequence conservation revealed that clusters of consensus motifs appeared in blocks of conserved regions (e.g. red dashed boxes in Figure 4.5), while motifs also occurred outside of conserved regions (e.g. grey dashed boxes in Figure 4.5). Damage score hits were also sometimes identified in regions with no conservation or consensus TFBS

(e.g. arrow head in Figure 4.5 a). In many cases, consensus binding motifs that were also evolutionarily conserved were not damage score hits (e.g. blue arrow head in Figure 4.5 a). However, when footprints aligned with deeply conserved regions that were also consensus TFBS (e.g. motifs marked with red stars in Figure 4.5), this suggested that these sequences were likely to be important for enhancer function.

4.2.5 Functional assays examining roles of upstream transcription factor binding sites in enhancer function

A combination of digital DNaseI footprinting, sequence conservation, and known TFBS was used to identify candidate upstream TFs regulating *Runx1* enhancers during EHT. Close examination of damage score hits in +23 enhancer revealed one hit that overlapped with a conserved Klf/Sp1 motif (GGGTGGG, Figure 4.5 a). This sequence produced a convincing DNaseI footprint in both HE and 416B cells (Figure 4.6 b). Quantification using shoulder-to-footprint ratio (SFR) showed footprint strength was similar in both cell types (HE SFR=1.34, 416B SFR=1.37), suggesting that this motif is likely to be bound by TFs in both cell types (Figure 4.6 b). Interestingly, Klf/SP1 motifs were not present in either +110 or +204 enhancers, suggesting that Klf/SP1 factors might be uniquely regulating +23 in HE and HP cells (Figure 4.6 a). A deeply conserved Cebp consensus binding site (TTGAGCAA) was seen in the +110 enhancer, but it was not a damage score hit in either cell type (red CEBP motif box in Figure 4.5 b). The requirement of Sasquatch to scan up to 7-mer motifs might have precluded the 8-mer Cebp motif from producing a DNaseI footprint. Interestingly, Cebp motifs were not seen in +23 enhancer, and only a weakly conserved motif was present in +204. By ChIP-seq, Cebpβ bound to the +110 enhancer in HP but not HE cells (Goode et al., 2016) (Figure 4.6 a, c). Together, this suggests that Cebp factors might be regulating the +110 enhancer in HP cells. Two damage score hits were seen at the +204 enhancer only in HE cells over two deeply conserved AP-1 motifs (TGACTCA, red stars in Figure 4.5 c). The motif produced DNaseI footprints in both cell types, but the footprint was 64.3% reduced in strength in 416B cells compared to HE (HE SFR=1.56, 416B SFR=1.20) (Figure 4.6 db). AP-1 motifs were not seen in the +23 enhancer, while two less conserved motifs were found in +110 (Figure 4.6 a, d). This suggested that AP-1 factors might be responsible for regulating +204 enhancer in HE cells specifically.

To examine whether these conserved TFBS within *Runx1* enhancers may be involved in regulating enhancer activity, I mutated the enhancers and tested them in luciferase-based enhancer-reporter assays. In total 4 Klf/SP1 binding sites were mutated in +23, 1 Cebp binding site was mutated in +110, and 2 AP-1 binding sites were mutated in +204 (solid red and red outline motifs in Figure 4.5, full length annotated enhancer sequences with mutations done are in Appendix Figures 7.5, 7.6, 7.7). In each case, enhancer activity significantly reduced compared to the wild type enhancer sequences when the newly identified TFBS was mutated, indicating a critical requirement for these specific motifs (Figure 4.7 a, one-way ANOVA, $F(1,11) = 87.5$, $p = 1.3\times10^{-16}$, $\eta^2 = 0.976$, *, adjusted $p < 0.001$, Tukey's post hoc test).

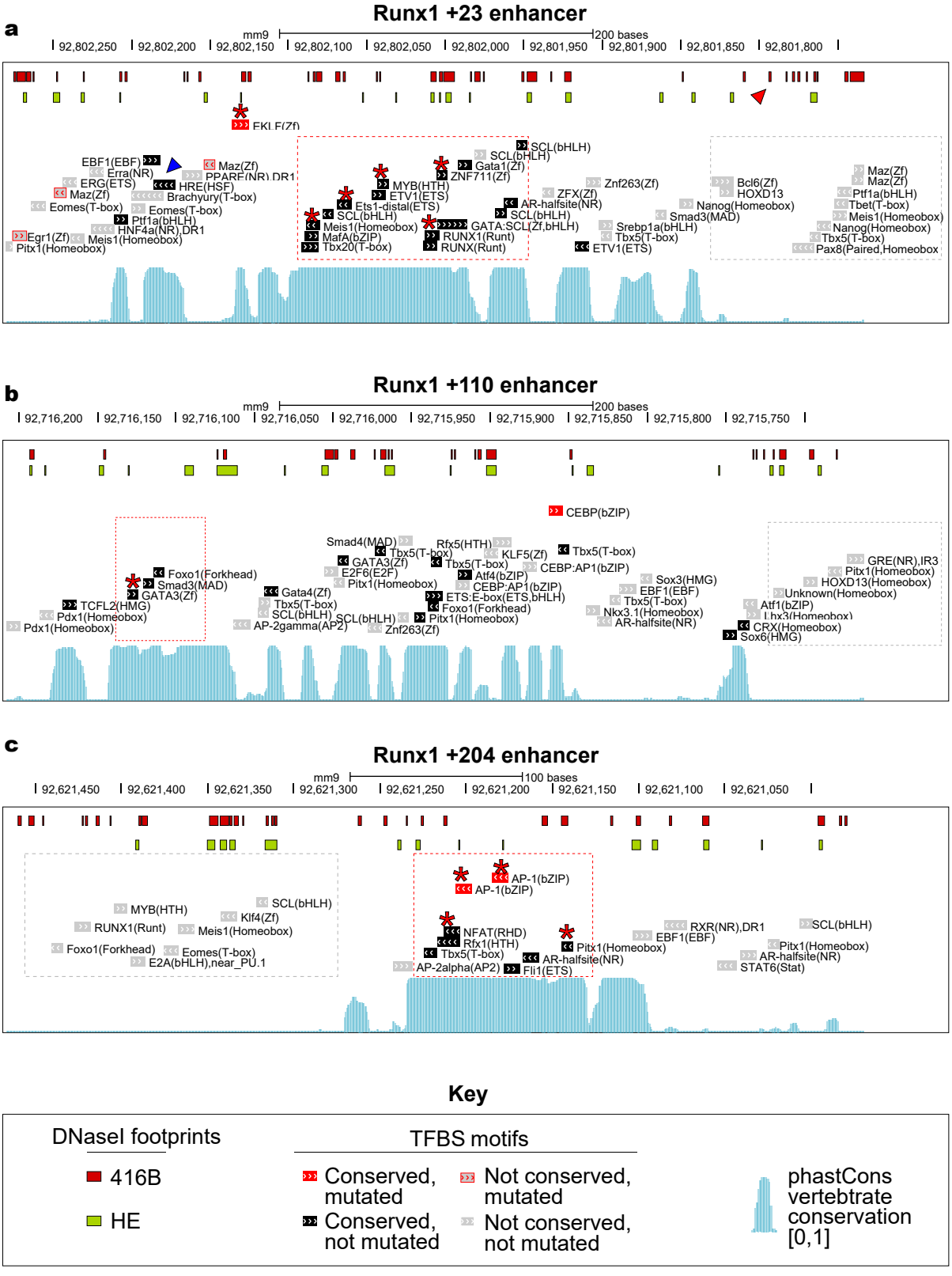


Figure 4.5 – DNaseI footprinting, evolutionary conservation, and consensus transcription factor binding motifs in *Runx1* enhancers *Legend continued on next page.*

Figure 4.5 – *Legend continued from previous page.* a-c) Evolutionarily conserved region comprising the +23 enhancer (Nottingham et al., 2007). DNaseI footprints identified in HE (green bars) and 416B (red bars) cells are shown. PhastCons vertebrate conservation (human, mouse, rat, chicken, and *Fugu rubripes*) is shown (Siepel et al., 2005). Consensus binding sites taken from HOMER (Heinz et al., 2010) that overlapping deeply conserved regions (phastCons 1.0) and non conserved regions are indicated as black and grey boxes with, respectively, with arrows indicating orientation. in a) A DNaseI footprint aligning to a deeply conserved Runx motif in the +23 enhancer is indicated by a red star. A DNaseI footprint that aligned to a non conserved region without any consensus motifs is indicated by a red arrowhead. A conserved consensus binding site that was not a DNaseI footprint is indicated with a blue arrow head. A cluster of conserved consensus binding sites are outlined with a red dashed box and a cluster of non-conserved binding sites are indicated by a grey dashed box. UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_DNaseI_ftpt`.

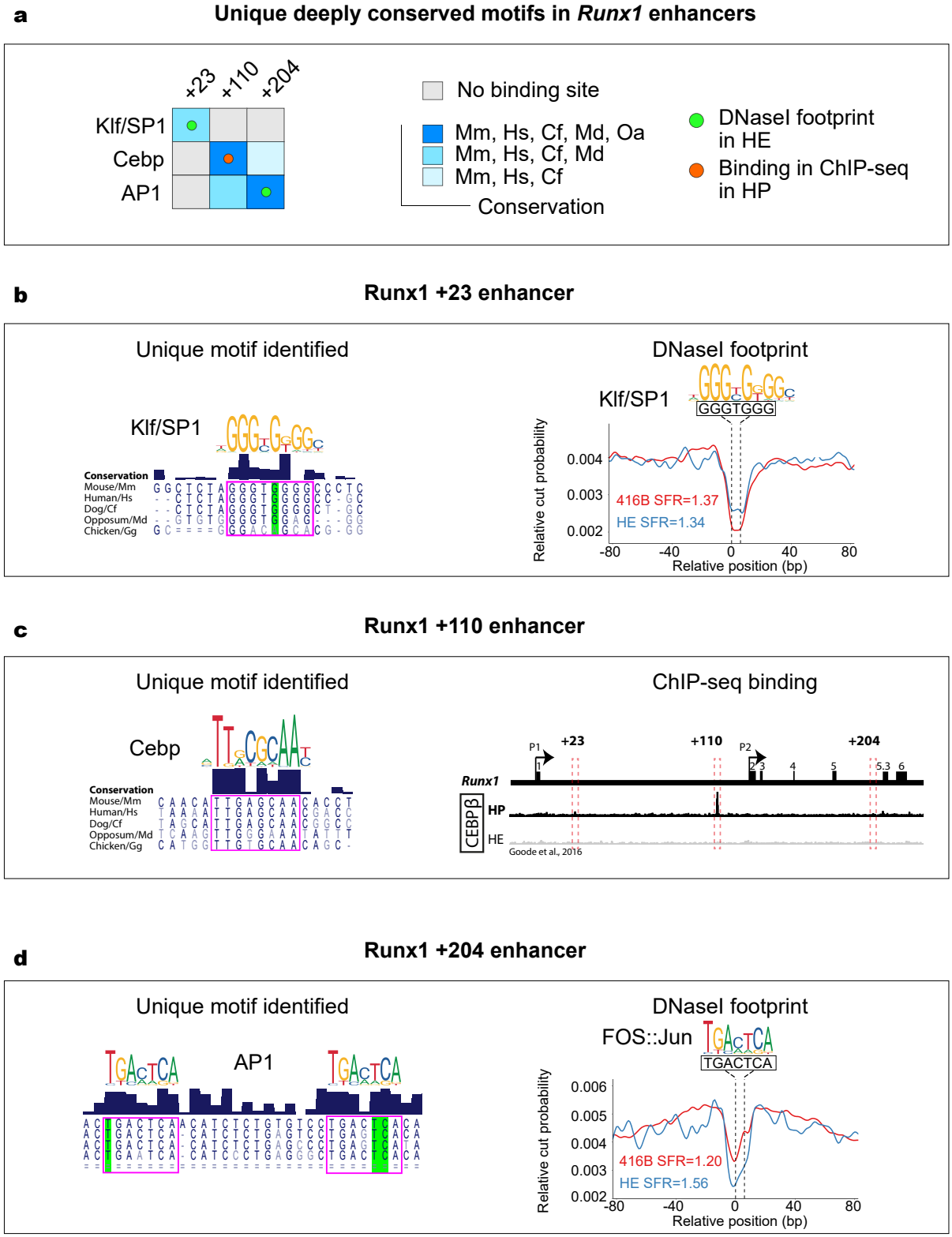


Figure 4.6 – Identification of unique deeply conserved transcription factor motifs in *Runx1* enhancers *Legend continued on next page.*

Figure 4.6 – *Legend continued from previous page.* a) Punnet square summarising conserved motifs that were unique between +23, +110, and +204 enhancers. Darker blue squares indicate more deeply conserved motifs, and grey squares indicates that there was no consensus motif identified. A green circle in a square indicates that a DNaseI footprint was identified in HE cells (Goode et al., 2016). An orange circle in a square indicates that binding was observed in ChIP-seq in HP cells (Goode et al., 2016). b) Conserved consensus motif for Klf/SP1 identified in +23 enhancer that was a DNaseI footprint in HE and 416B cells. The green base in the conservation plot yielded a significant 'damage score' upon *in silico* mutation using Sasquatch (Schwessinger et al., 2017). Shoulder-to-footprint ratio (SFR) is a measure of footprint strength (Schwessinger et al., 2017). c) Cepbp site that was deeply conserved in +110 enhancer only. Publicly available ChIP-seq for *Cebpβ* in HE and HP cells (Goode et al., 2016) indicated a binding peak only at the +110 enhancer and only in HP cells. d) A pair of deeply conserved AP-1 binding sites identified in +204 enhancer that produced a significant DNaseI footprint in HE cells and to a lesser extent in 416B cells.

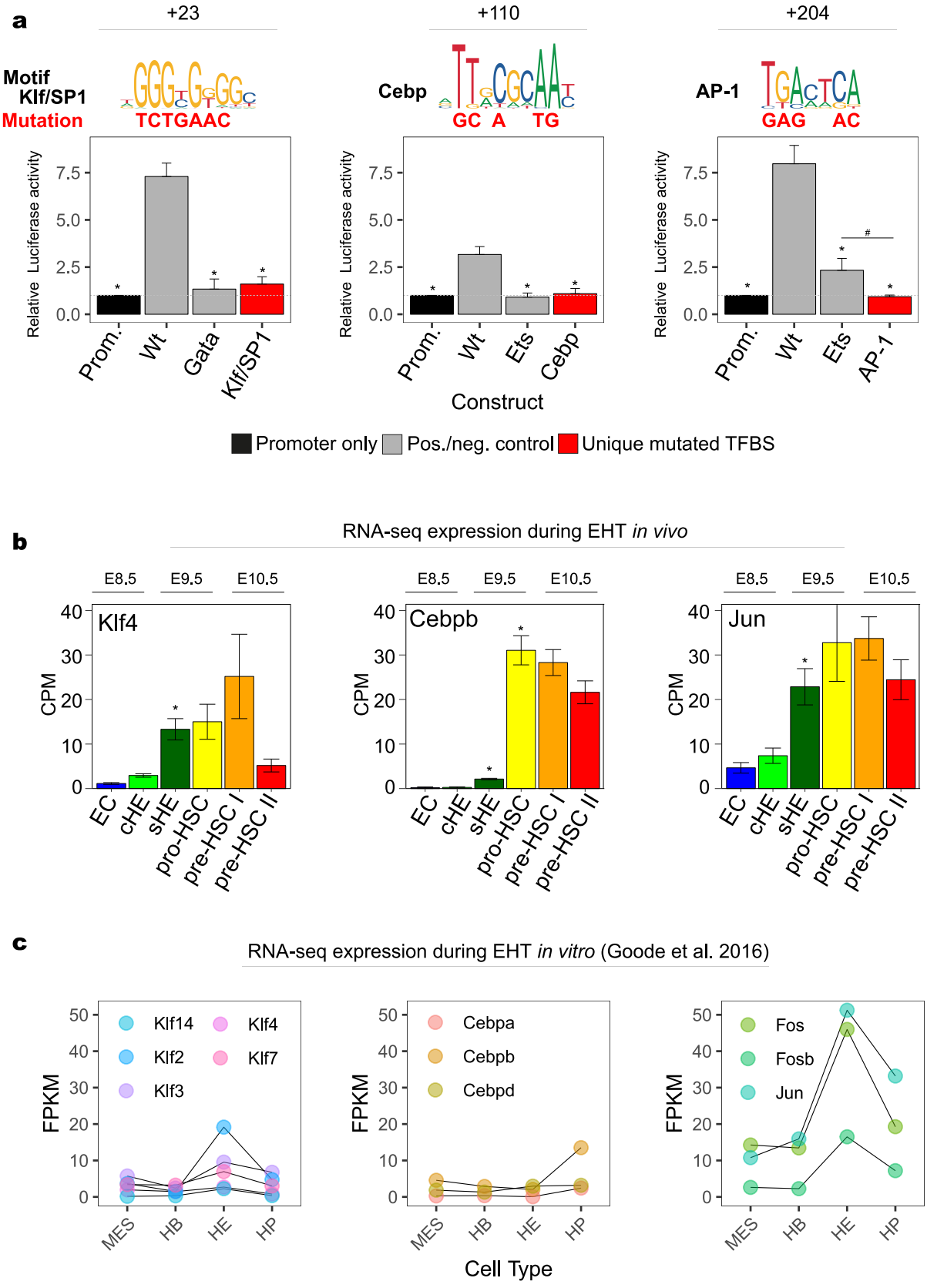


Figure 4.7 – Exploring possible roles for novel upstream regulators of *Runx1* enhancers
Legend continued on next page.

Figure 4.7 – *Legend continued from previous page.* a) Consensus binding motifs and mutations that were done to enhancer sequences that were then tested in luciferase-based enhancer-reporter assays in 416B cells. Positive control was the wild type (Wt) enhancer sequence, and negative control was a mutation that had previously been shown to abrogate the activity of each enhancer (Schütte et al., 2016). The red bar indicates the activity of enhancer after the novel binding site was mutated as shown. One-way ANOVA, $F(1,11) = 87.5$, $p = 1.3\times10^{-16}$, $\eta^2 = 0.976$, *, adjusted $p < 0.001$, #, adjusted $p < 0.05$ Tukey's post hoc test. b) RNA-seq expression data of selected upstream TFs in sorted populations of mouse cells undergoing EHT *in vivo* (Joe Harman, Lucas Greder, Gemma Swiers, unpublished results). *, FDR < 0.05. c) Expression of selected upstream TFs by RNA-seq in sorted populations of mESC-derived cell populations undergoing *in vitro* EHT (Goode et al., 2016).

To support a role for these TFBS in *Runx1* transcriptional regulation during EHT, RNA-seq expression data in populations undergoing EHT *in vivo* in mouse embryos (de Bruijn lab unpublished results), and *in vitro* in mESC-derived populations (Goode et al., 2016) was analysed. Several TFs capable of binding the motifs identified above (Figure 4.6) were stage-specifically and significantly upregulated during EHT (Figure 4.7 b and c). Several *Klf* family genes including *Klf4* were upregulated in HE cells (Figure 4.7, b, c *, FDR < 0.05), where +23 was previously shown to drive reporter gene expression (Nottingham et al. 2007; Swiers et al. 2013a, Figure 1.4 a). Expression of several *Cebp* genes including *Cebpβ* increased from HE to HP cells (Figure 4.7, b, c *, FDR < 0.05), where the +110 enhancer was previously shown to be active (de Bruijn lab unpublished observations, Figure 1.4 a). AP-1 family TFs including *Jun*, *Fos*, and *Fosb* were also specifically upregulated in HE cells (Figure 4.7, b, c *, FDR < 0.1), where the +204 enhancer-reporter transgene was expressed (de Bruijn lab unpublished observations, Figure 1.4 a). Together, the expression of these TFs across EHT were correlated with the stages at which the enhancers were previously shown to be active during EHT *in vitro* (Figure 4.4) and *in vivo* (Figure 1.4 a, Nottingham et al. 2007 and de Bruijn lab unpublished results). Despite many similarities between the upstream TFBS regulating *Runx1* enhancers, differences also exist, suggesting that they might play different roles in regulating *Runx1* expression at discrete stages of EHT. However, it remains unclear which *Runx1* enhancers in the endogenous locus are functionally required for its transcription at specific stage of EHT.

4.2.6 CRISPR/Cas9 nickase can be used to examine functional requirements of *Runx1* enhancers

To investigate functional requirements from Runx1 enhancers, we used a dual paired (4x sgRNA) CRISPR/Cas9^{D10A} nickasestrategy to delete +23, +110, and +204 Runx1 enhancers in 23C-RV mESCs. Intended deletions ranged from 1 to 1.5 kb (Figure 4.8 c). Individual clones were analysed for the desired genotype using S-R PCR (Figure 7.8 b). Out of 445 clones analysed (110 targeting the +23 enhancer, 169 targeting the +110 enhancer, 166 targeting the +204 enhancer), an average of 35% and 20% of the total isolated clones for each of the three targeted sites appeared to be homozygous knock-out or wild type, respectively (Appendix Figure 7.8 c, d). Several clones with unique alleles harbouring deletions of expected size (EDs, spanning less than 25 bp from expected sgRNA cut sites) were mapped using Sanger sequencing of PCR products (Figure 4.8, c, grey lines). Sanger sequencing often generated a single sequencing trace, indicative either of an iso-allelic HR event (both alleles carrying the same deletion) or loss of a primer binding site, leading to failure to amplify one of the alleles (allelic drop-outs) (Figure 7.8 e). Inconsistent results obtained from haematopoietic differentiation of multiple enhancer-deleted clones (data not shown) unfortunately suggested that the latter might be the case.

4.2.7 Larger deletions occur at a high frequency after CRISPR/Cas9 deletions

To investigate the genotypes of clones with possible allelic drop-outs we performed PCR screening using medium-range (M-R) PCR, with primers located >600 bp away

from the sgRNA cut sites (Figure 4.8 b, c). We found several clones that harboured a LD that was not detected using S-R primers (Figure 4.8 b, c). Indeed, multiple deleted alleles not detected using S-R PCR were observed in a pool of targeted and selected mESCs (Figure 4.8 b). Out of a total of 84 clones that were assigned a homozygous knock-out or wild type genotype based on S-R PCR, 13 (15%) harboured a LD on one allele only detected by M-R PCR (Figure 4.8 b, d). Five clones were further analysed using Sanger sequencing which confirmed *bona fide* LDs on one allele (Figure 4.8 c). The deletions spanned 300-600 bp from either of the sgRNA cut sites at each of the three enhancers. Interestingly, one clone contained a secondary deletion upstream from the original cut site that removed one of the S-R PCR primer binding sites (Figure 4.8 c, *, mid panel). Longer-range (L-R) PCR amplifying 5.5 kb fragments revealed even larger LDs spanning up to 2.7 kb away from sgRNA cut sites in pools of selected cells and isolated clones (Figure 4.8 c, light green dashed lines).

Importantly, PCR and sequencing based approaches may be impacted by biases including over amplification of shorter alleles containing LDs. To accurately quantify LDs in a $Cas9^{D10A}$ nickase-targeted cell population without amplification bias, we utilised digital droplet PCR (ddPCR). Amplicons spaced at 500 bp, 1 kb, and 3 kb intervals away from 4x sgRNA cut sites revealed that relative target DNA concentration was significantly reduced in pools of selected cells compared to non-targeting controls (Figure 4.8 e). Similar LDs were also found when targeting five CTCF sites at the *Runx1* locus (P1 CTCF, P2 CTCF, and downstream [ds] CTCFs 1 to 3, Appendix Figure 7.9), and *Prickle2* exon 6 using one or two sgRNAs (1x or 2x sgRNAs, Appendix Figure 7.10). Our collaborators (Lydia Teboul, Adam Caulder, Alasdair Allan, Gemma Codner) also identified several LDs after gene editing in mice (Appendix Figure 7.11). These findings reveal that LDs are readily detectable in pools of targeted cells, isolated clones, and in animals, irrespective of genomic site or genome editing methodology used. Moreover, LDs are still detectable even without biases due to LR PCR primer position or design.

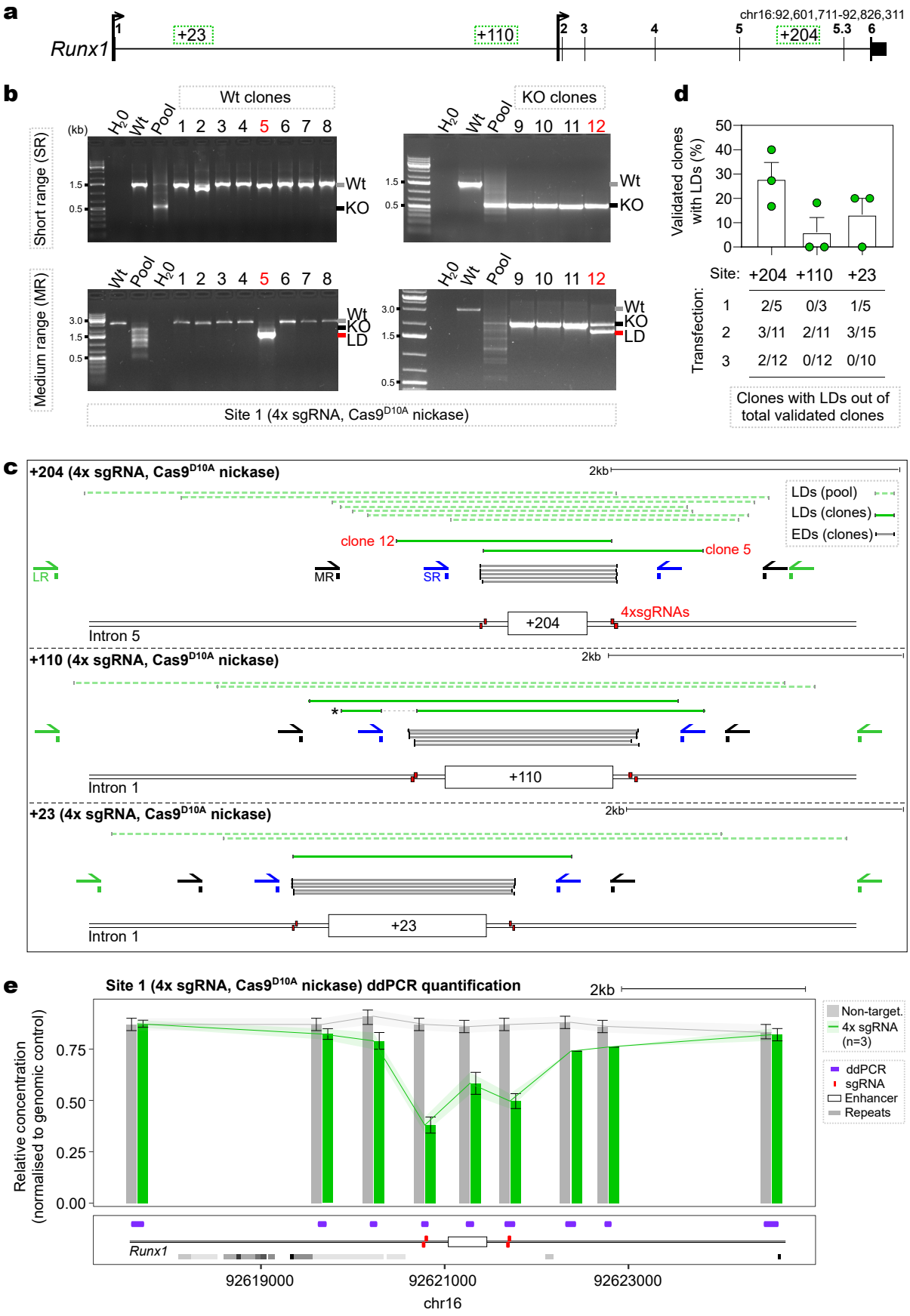


Figure 4.8 – Cas9^{D10A} nickasecan be used to examine *Runx1* enhancer requirements

Legend continued on next page.

Figure 4.8 – *Legend continued from previous page.* a) Locus map of the *Runx1* gene showing the positions of evolutionarily conserved enhancers (+23, +110, +204) that were targeted in E14-TG2a-RV mESCs using $Cas9^{D10A}$ nickase. b) Example gel images from one experiment targeting +204. Gel images show PCR amplification from gDNA of isolated wild type (wt) clones (left hand gels) and knock-out (KO) clones (right hand gels) with SR primers (top gels) and MR primers (bottom gels). Wt next to the gel image indicates the size of the wild type allele, KO indicates the size of alleles harbouring the expected deletion, and LD indicates the size of alleles in clones identified as harbouring LDs. c) Schematic showing the positions of short-range PCR primers (SR, blue), medium-range PCR primers (MR, black), longer-range PCR primers (LR, green), sgRNAs (red boxes), and LDs isolated from clones (dark green lines) and pools of cells (light green dashed lines) at each of the three enhancers. The allele marked with a star contained a secondary deletion at +110 distal to the primary cut site that destroyed a primer binding site. d) Quantification of clone frequencies with homozygous wild type or knock-out genotypes by short-range PCR (validated clones) that were later found to contain a LD on another allele by medium-range PCR. Quantification of clone numbers for each transfection that were homozygous knock-out or wild type and contained a LD on the other allele (n=3 independent transfections per site, each dot is one independent experiment). e) ddPCR quantification of deletions across a 7 kb window centred on Enhancer 1. Each bar represents the mean relative concentration of the target DNA sequence (+/- 95% confidence interval). mESCs were targeted with 4x sgRNA (blue bars, n=3) and a non-targeting control (grey bar). Alasdair Allan performed ddPCR on genomic DNA prepared by me.

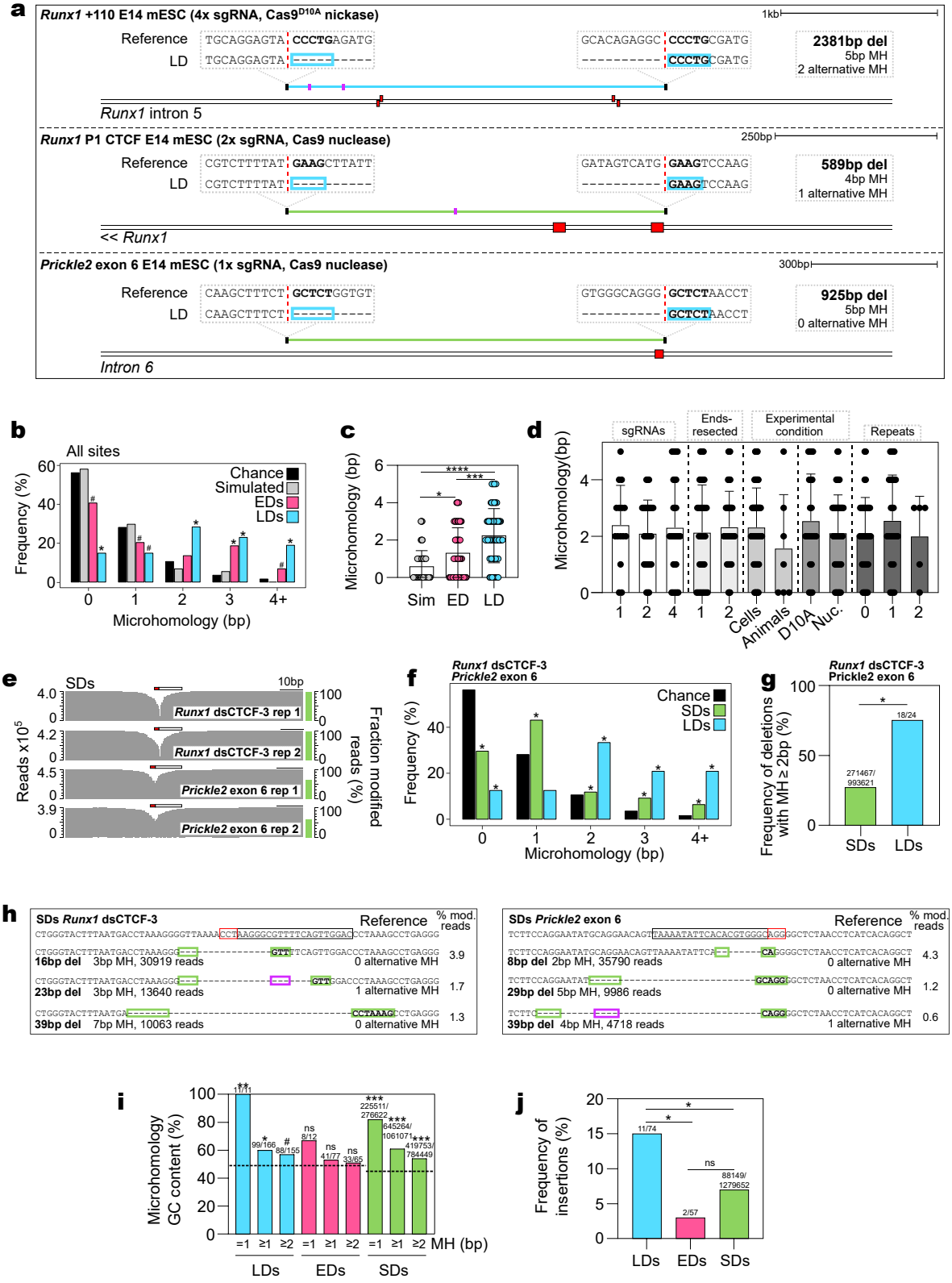


Figure 4.9 – Microhomologies consistent with MMEJ are prevalent at Cas9-induced larger deletions. *Legend continued on next page.*

Figure 4.9 – *Legend continued from previous page.* a) Examples of LDs (blue and green lines) with microhomologies and corresponding reference sequences shown (mm9). Sequences outlined with blue boxes represent microhomologies. Red dashed vertical lines represent the exact breakpoint junctions in the repaired alleles and sgRNAs are shown (red boxes). Total deletion size, microhomology amount, and number of alternative (more proximal) microhomologies are shown (pink lines in deleted sequence). b) Frequency distribution histogram of microhomologies at 74 LD breakpoint junctions (LDs) across 16 sites, 59 expected size deletions across 16 sites (EDs), 74 simulated deletions (Simulated), and the chance expectation of finding at two locations a k-mer of a given length (Chance) (χ^2 test, *, $p < 6^{-7}$, #, $p < 0.02$). c) Microhomology at 74 LDs compared to 59 EDs, and 74 simulated deletions (Sim) (two-tailed Kruskal-Wallis test, ****, $p < 0.0001$, ***, p=0.0007, *, p=0.0105). d) Comparison of microhomology at LDs generated with 1, 2 or 4 sgRNAs, with ends resected in one or two directions, generated under different experimental conditions, or intersecting with 0, 1, or 2 repeat elements (two-tailed Kruskal-Wallis test, $p > 0.9999$). e) Short-amplicon sequencing from pools mESCs targeted using one sgRNA. Positions of protospacer (black outlined bar) and protospacer adjacent motif (PAM, red bar) are shown. Fraction of modified reads and read counts are shown. Protospacer (black outlined bar) and PAM (red outlined bar) are indicated. f) Microhomology quantification in 24 larger deletions (LDs) and all shorter deletions (SDs) mapped at *Runx1* dsCTCF-3 and *Prickle2* exon 6 compared to the chance expectation of finding a k-mer of a given length (χ^2 test, *, $p < 0.0003$). g) Quantification of deletions containing microhomology $\leqslant 2bp$ in all SDs and LDs generated at *Runx1* dsCTCF-3 and *Prickle2* exon 6 using one sgRNA (χ^2 test, *, $p=5.4^{-7}$) h) Reference sequence and Cas9-induced deletion alleles containing significant microhomologies at their breakpoints. The total number of reads and the percentage of modified reads is shown. Protospacer (black outlined bar) and PAM (red outlined bar) are indicated. Short microhomologies that abut the deletion are outlined with green boxes and alternative microhomologies located within the deleted region are outlined with pink boxes. i) Quantification of microhomology GC base pair content in microhomologies of different lengths at all LDs and expected size deletions (EDs) and SDs at *Runx1* dsCTCF-3 and *Prickle2* exon 6. The expected background GC base pair content is shown as a black dashed line. (χ^2 test, ns, $p > 0.2$, #, p=0.059, *, $p < 0.01$, **, $p < 0.001$, ***, $p < 10^{-10}$). j) The number of total LDs, total EDs, and SDs at *Runx1* dsCTCF-3 and *Prickle2* exon 6 containing a short insertion (χ^2 test, *, $p < 0.003$, ns, p=0.2395).

4.2.8 Microhomologies consistent with MMEJ are prevalent at larger deletions

Next, we explored whether the new DNA sequences that were created after DNA breaks could inform on potential DNA repair mechanisms associated with LDs. Microhomologies of 2-5 bp in length were found at 52 out of 74 (70%) of the LD breakpoint junctions we identified (Figure 4.9 a). Homologous base pair scoring identified significantly more microhomology at LDs compared to simulated LDs, microhomology expected by chance for a k-mer of a given length, and microhomology found at EDs (Figure 4.9 b and c). LDs contained microhomologies irrespective of whether LDs were generated with single or multiple sgRNAs, exhibited DNA end-resection at 1 or 2 adjacent DSBs, were generated *in vitro* or *in vivo*, by $Cas9^{D10A}$ nickase or Cas9 nuclease (Figure 4.9 d). There was no difference between the length or frequency of microhomologies found at LD breakpoints associated with 0, 1, or 2 annotated repeats (Figure 4.9 d), nor were annotated repeats enriched at LD breakpoint junctions (Appendix Figure 7.12). In addition to our own data, I analysed 69 Cas9-induced LDs from the literature that were previously generated by single or pairs of sgRNAs (Mianne et al., 2017; Ma et al., 2014; Zhou et al., 2014; Wang et al., 2013; Parikh et al., 2015; Zhang et al., 2015; Adikusuma et al., 2018). These 69 distinct LDs also contained a significant over-representation of microhomologies compared to the chance expectation or simulated deletions (Appendix Figure 7.13). Altogether, these data suggest that microhomology-mediated end-joining (MMEJ) is active during the repair of LD alleles, as MMEJ depends on short (<20 bp) microhomologies that are shared between both breakpoints, with some tolerance for mismatches (McVey and Lee, 2008).

MMEJ was previously implicated in the repair of Cas9-induced DSBs at shorter deletion alleles (SDs) of less than 60 bp (van Overbeek et al., 2016; Bae et al., 2014; Shen et al., 2018; Ata et al., 2018; Allen et al., 2018; Taheri-Ghahfarokhi et al., 2018; Chakrabarti et al., 2019; Kim et al., 2018; Brinkman et al., 2018). To directly compare the prevalence of microhomologies at Cas9-induced SDs with LDs, I performed short-amplicon deep sequencing after targeting two different regions with 1x sgRNA (*Runx1* dsCTCF-3 and exon 6 of *Prickle2*, Figure 4.9 e). LDs at both sites were significantly enriched for microhomologies compared to SDs quantified at the same sites (Figure 4.9 f, g). Still, microhomologies were significantly over-represented at SDs compared to the chance expectation of two sequences containing a k-mer of a given length (Figure 4.9 f, h). MMEJ has previously been shown to favour thermostable microhomologies with elevated GC content (Glover et al., 2011; Shen et al., 2018). Microhomologies at all Cas9-induced LDs and SDs at *Runx1* dsCTCF-3 and *Prickle2* were both significantly enriched for GC base pairing compared to background (Figure 4.9 i), while microhomologies across all EDs were observably but not significantly enriched (Figure 4.9 i). Interestingly, GC bases were always the most enriched in microhomologies of 1 bp, compared to longer microhomologies (Figure 4.9 i). MMEJ is also known to frequently generate small non-templated insertions (Yousefzadeh et al., 2014; Sfeir and Symington, 2015). In line with this, Cas9-induced LDs were enriched for small insertions compared to EDs and SDs (Figure 4.9 j). Collectively these data show that most Cas9-induced LDs contain microhomologies consistent with MMEJ at their breakpoints.

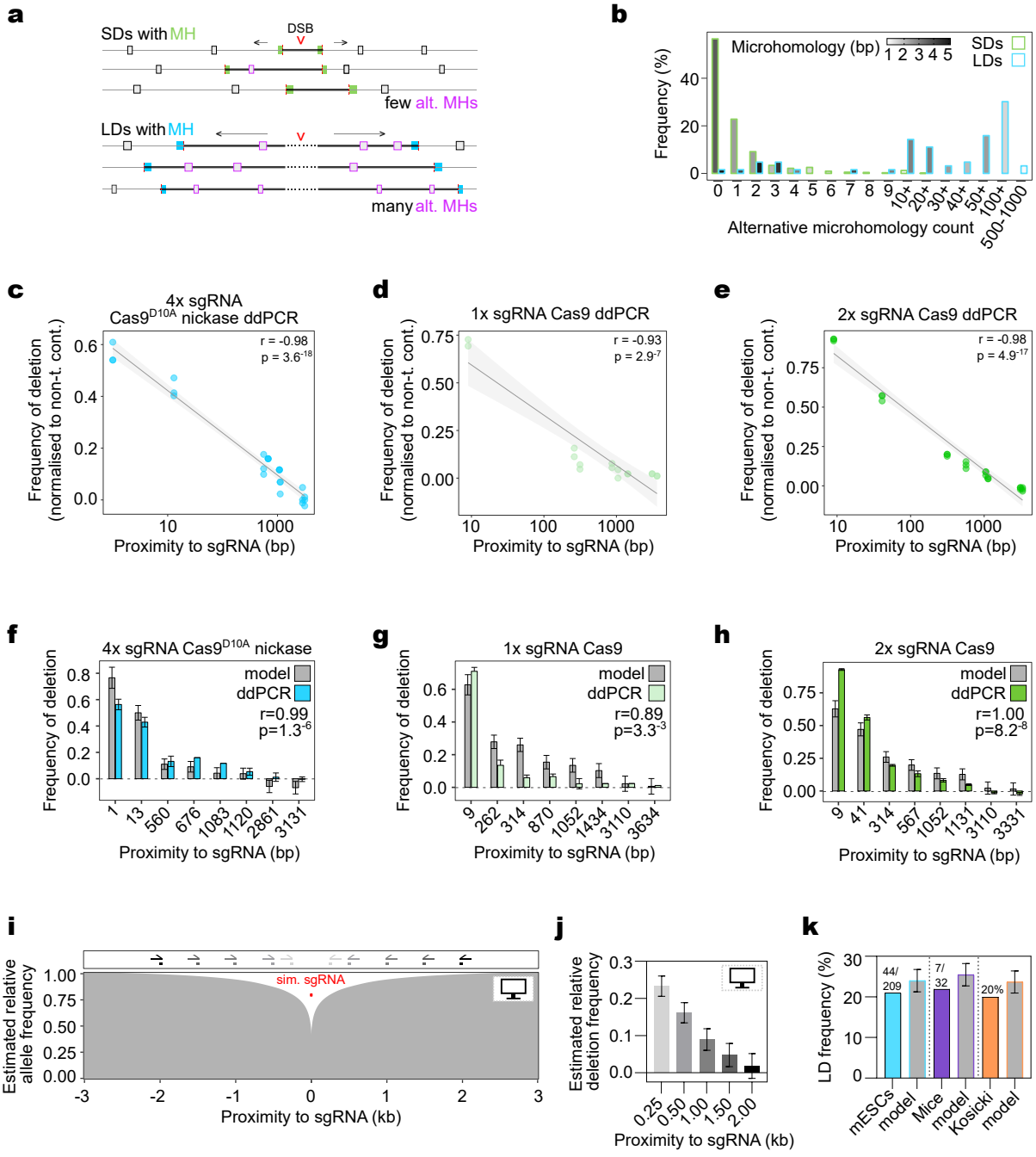


Figure 4.10 – Larger deletion breakpoints do not occur at proximal microhomology sequences but are dependent on proximity to sgRNAs. a) Schematic representation of LDs and SDs undergoing end-resection and bypassing alternative more proximal microhomologies during DNA repair. Shorter deletion microhomologies are shown in green, larger deletion microhomologies are shown in blue, and alternative microhomologies are indicated by pink boxes. The sequence included in the deletion is shown as a bold black line. b) Quantification of alternative microhomologies that were found in the deleted sequences in SDs at *Runx1* dsCTCF-3 and *Prickle2* exon 6, and across all LDs with microhomologies at their breakpoints. The colour gradient represents the mean microhomology score of all deletions within each bin. *Legend continued on next page.*

Figure 4.10 – *Legend continued from previous page.* c-e) Correlation between frequency of deletion determined by ddPCR and sgRNA proximity. Pearson correlation r and p values are indicated and a linear regression with 95% confidence interval is shown. f-h) Deletion frequencies of real ddPCR data and model estimates with Pearson correlation r and p values indicated. i) Model estimate of deletion frequency over a 6 kb window around a simulated sgRNA cut site with simulated PCR primers indicated as grey to black half arrows above the plot. j) Relative predicted deletion frequencies at each of the simulated primer sites with 95% confidence intervals indicated. k) Comparison between estimated and empirically determined deletion frequencies in two of our own independent data sets and one recent experiment reported in the literature (Kosicki et al., 2018). Alasdair Allan performed ddPCR on genomic DNA prepared by me.

4.2.9 Larger deletions cannot be predicted by microhomology sequences and are dependent on proximity to cut sites

Recent work has suggested that DNA repair outcomes are predictable at Cas9-induced DSBs based on the presence of microhomologies in cut site-proximal DNA sequences (Bae et al., 2014; Shen et al., 2018; Ata et al., 2018; Allen et al., 2018; Taheri-Ghahfarokhi et al., 2018). We asked whether the distribution of LDs was similarly dependent on the proximity of microhomologies to cut sites. At LDs, deletion size was independent of microhomology length, unlike at SDs (Appendix Figure 7.14). For all but one LD, the intervening sequence between deletion ends and sgRNA cut sites contained several alternative (more proximal) microhomologies that were bypassed during repair (median=49, Figure 4.10 a and b). In contrast, microhomologies used for repair at SDs were predominantly (but not exclusively) the most proximal to the cut site (Figures 4.10 a and b, 4.9 h). The number of alternative microhomologies present in the deleted sequence was dependent on deletion length and microhomology length (Appendix Figure 7.14 b-d), reflecting the random distribution of microhomology sequences throughout the genome. Together this indicates that in contrast to SDs, LDs are not repaired to the closest microhomology.

Given the fact that LD sizes are independent of cut site proximal microhomology sequences, we examined what other factors might influence LD formation. At the population level, the distribution of deletion sizes as inferred from ddPCR was negatively correlated with proximity to sgRNA target sites (Figure 4.10 c-e). I modelled this relationship using multiple linear regression and found that over 80% of the variance in the distribution of deletion sizes depended on proximity to sgRNAs and sgRNA cutting efficiency determined by ddPCR (Figure 4.10 f-h, Figure 7.15, adjusted R^2=0.8275, $p < 2^{-16}$). Interestingly, the model based on empirical ddPCR measurements estimated that in general 22 +/- 3% of alleles were deleted 250 bp from sgRNA target sites (Figure 4.10 i and j). This agrees with the 21% (44 out of 209) of our isolated mESC clones that harboured LDs abolishing S-R PCR primers (Figures 4.8 d, 7.9 d, 4.10 k, mean sgRNA proximity=243 bp). Furthermore, S-R primer binding sites with a mean sgRNA proximity of 211 bp were abolished in 22% (7 out of 32) mouse projects (Figure 7.11, Figure 4.10 k). A recent study also found that LDs >250 bp occurred in up to 20% of alleles (Figure 4.10 k, Kosicki et al. 2018; Thomas et al. 2019). The fact that LDs occur in a predictable manner in populations of cells allows the rational design of effective genotyping strategies to detect them.

4.2.10 Identification of homozygous *Runx1* enhancer knock-out clones using an optimised genotyping protocol

Using insights gained from analysing large numbers of Cas9-induced LDs, I sought to identify mESC clones with homozygous *Runx1* enhancer deletions that could be used to determine functional enhancer requirements. I used a combined genotyping approach of M-R (3 kb) and L-R (5.5kb) PCRs and ddPCR copy counting. Only clones that were shown by S-R genotyping to harbour homozygous deletions at +23, +110 or +204 enhancers individually were examined. Each of the clones exhibited a single Sanger sequencing trace, indicative either of identical homozygous deleted alleles or allelic drop-out (Figures 4.11, 4.12, 4.13 a). Several of these clones harboured

LDs that were only detected upon M-R or L-R PCR (Figures 4.11, 4.12, 4.13 b and c, red lines next to gels). Copy counting by ddPCR revealed that some of these LDs also deleted the enhancer (Figure 4.13 c and e, clone Δ204-D7, Appendix Figure 7.18 green star), while others did not (Figure 4.11 c and e, clone Δ23-D12, Appendix Figure 7.16 red stars). Interestingly, ddPCR copy counting identified clones harbouring alleles that were not resolved by L-R PCR. For example, only PCR products indicative of a knock-out allele were seen for clone Δ110-B3 by L-R PCR, but by ddPCR the +110 enhancer was shown to still be present (Figure 4.12 c and e, clone Δ110-B3, Appendix Figure 7.17 red star). Reinsertion elsewhere of the cut out +110 enhancer may have occurred in this clone, which has been observed before (Boroviak et al., 2017). Clone Δ204-B3 appeared to harbour a deletion of the +204 enhancer, but by ddPCR harboured a duplication and three copies of the region downstream of the sgRNAs (Figure 4.13 c and e, clone Δ204-B3, Appendix Figure 7.18 blue stars). This further highlights the importance of combining L-R PCR with copy counting to fully characterise enhancer deletions. After rigorous genotyping, clones Δ110-E4.1, Δ204-D7, and Δ204-G8.1 were shown to likely harbour homozygous enhancer deletions with an otherwise intact locus and would allow functional requirements for these *Runx1* enhancers to be investigated.

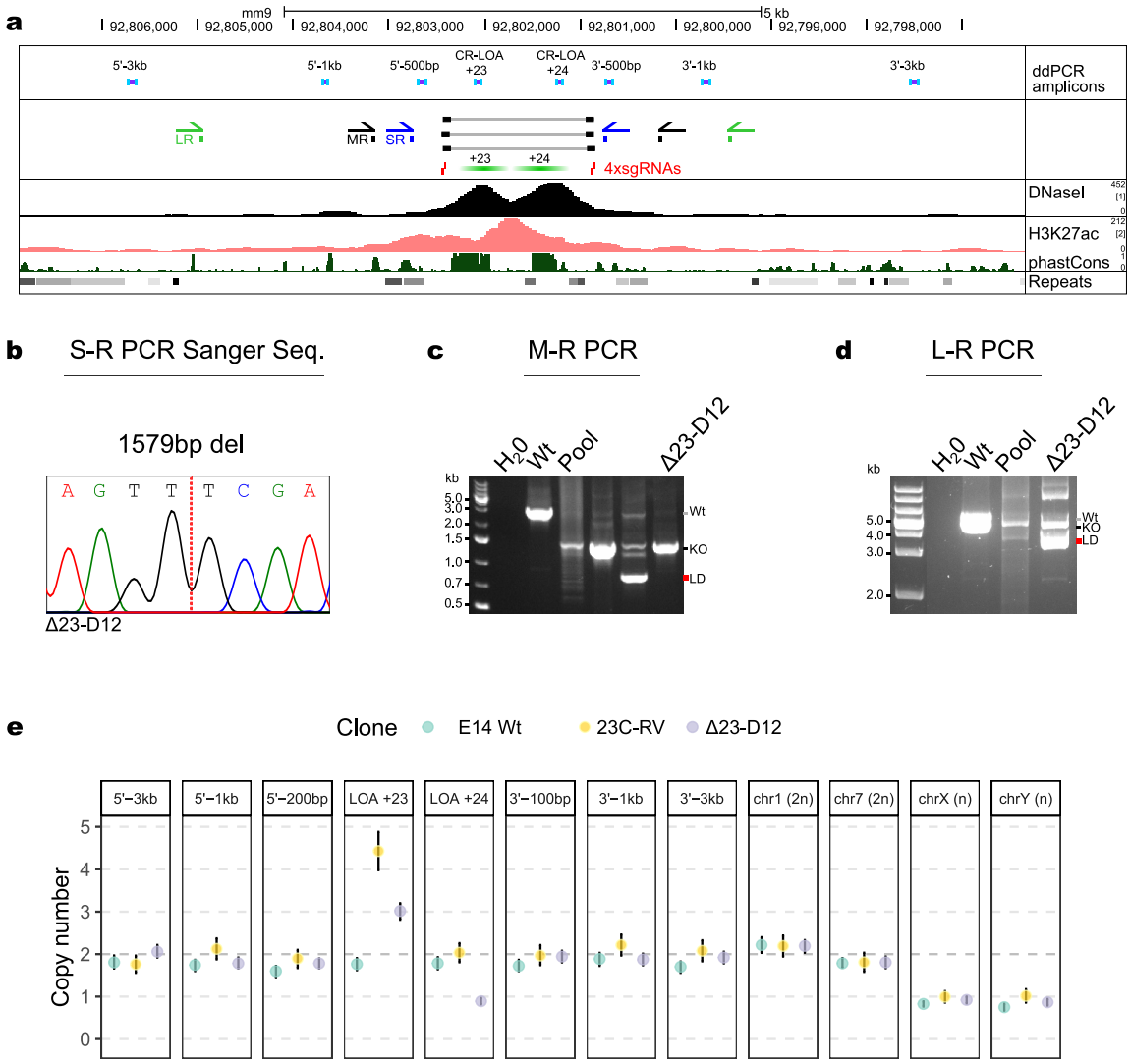


Figure 4.11 – Validation of *Runx1* +23 enhancer deletion clones. a) Locus map of targeting strategy used to delete *Runx1* enhancers +23/+24. Locations of 4x sgRNAs used alongside $Cas9^{D10A}$ nickaseare shown as red boxes, primers used in short-range (S-R), medium-range (M-R), and longer-range (L-R) PCR are indicated as half arrows. Digital droplet PCR (ddPCR) amplicons spaced over the enhancer region are shown as purple boxes with blue ends. Deletions mapped using S-R PCR and Sanger sequencing are indicated as grey boxes with black ends. The locations of +23 and +24 enhancers as defined by Schütte et al. 2016 are indicated as green boxes. DNaseI-seq ([1] Vierstra et al. 2014) and H3K27ac ChIP-seq ([2] Schütte et al. 2016) shown were previously generated in 416B cells. PhastCons vertebrate conservation and repeats are indicated. b) S-R PCR Sanger sequencing trace from clone Δ23-D12 that was shown to harbour a 1579 bp deletion (a red dashed line indicates the position of the deletion). This clone contained only a single sequencing trace potentially indicating an allelic drop-out. c) M-R PCR gel image for three clones including Δ23-D12 that were shown to harbour a knock-out for +23/+24 enhancers by S-R PCR and a single Sanger sequencing trace. Wt and a grey bar indicates the size of the wild type allele, KO and a black bar indicates the size of the knock-out allele, LD and a red line indicates a LD allele identified in the clones. *Legend continued on next page.*

Figure 4.11 – *Legend continued from previous page.* d) L-R PCR from clone Δ23-D12 with identical labelling to c. e) ddPCR genotyping over the +23/+24 enhancer region from E14 wild type, 23C-RV, and clone Δ23-D12. ddPCR droplets used to calculate relative copy numbers are shown in Appendix Figure 7.16. Relative copy numbers indicated were calculated relative to an internal control amplicon in each well and to two different diploid chromosomes (chr1 and chr7) that were not targeted using $Cas9^{D10A}$ nickase. Amplicons targeting chrX and chrY were included as single copy controls. E14-Wt cells were used as a diploid control (relative copy number 2) across all expected amplicons and a haploid control (relative copy number 1) for the sex chromosomes. 23C-RV cells were included as a 4n control as they were previously targeted with a +23-Cherry enhancer reporter construct and are homozygous for this knock-in (so relative copy number = 4 in total).

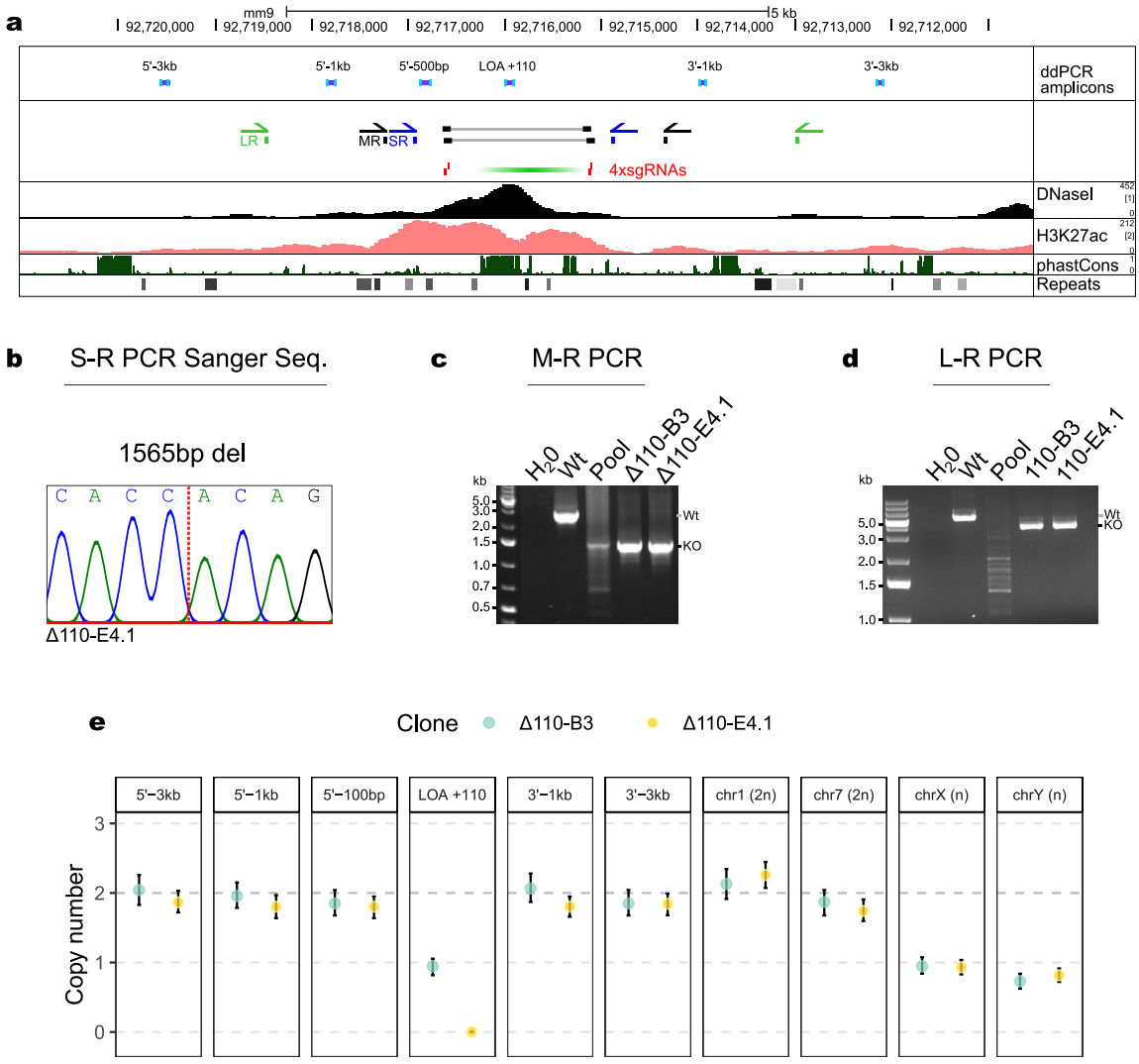


Figure 4.12 – Validation of *Runx1* +110 enhancer deletion clones. a) Locus map of targeting strategy used to delete *Runx1* enhancer +110. Locations of 4x sgRNAs used alongside $Cas9^{D10A}$ nickaseare shown as red boxes, primers used in short-range (S-R), medium-range (M-R), and longer-range (L-R) PCR are indicated as half arrows. Digital droplet PCR (ddPCR) amplicons spaced over the enhancer region are shown as purple boxes with blue ends. Deletions mapped using S-R PCR and Sanger sequencing are indicated as grey boxes with black ends. The location of +110 enhancer as defined by Schütte et al. 2016 are indicated as green boxes. DNaseI-seq ([1] Vierstra et al. 2014) and H3K27ac ChIP-seq ([2] Schütte et al. 2016) shown were previously generated in 416B cells. PhastCons vertebrate conservation and repeats are indicated. b) S-R PCR Sanger sequencing trace from clone Δ110-E4.1 that was shown to harbour a 1565 bp deletion (a red dashed line indicates the position of the deletion). This clone contained only a single sequencing trace potentially indicating an allelic drop-out. c) M-R PCR gel image for Δ110-B3 and Δ110-E4.1 that were shown to harbour a knock-out for +110 enhancer by S-R PCR and a single Sanger sequencing trace. Wt and a grey bar indicates the size of the wild type allele, KO and a black bar indicates the size of the knock-out allele. *Legend continued on next page.*

Figure 4.12 – *Legend continued from previous page.* d) L-R PCR from clones Δ110-B3 and Δ110-E4.1 with identical labelling to c. e) ddPCR genotyping over the +110 enhancer region from Δ110-B3 and Δ110-E4.1. ddPCR droplets used to calculate relative copy numbers are shown in Appendix Figure 7.17. Relative copy numbers indicated were calculated relative to a negative control amplicon in each well, and to two different diploid chromosomes (chr1 and chr7) that were not targeted using $Cas9^{D10A}$ nickase. Amplicons targeting chrX and chrY were included as single copy controls.

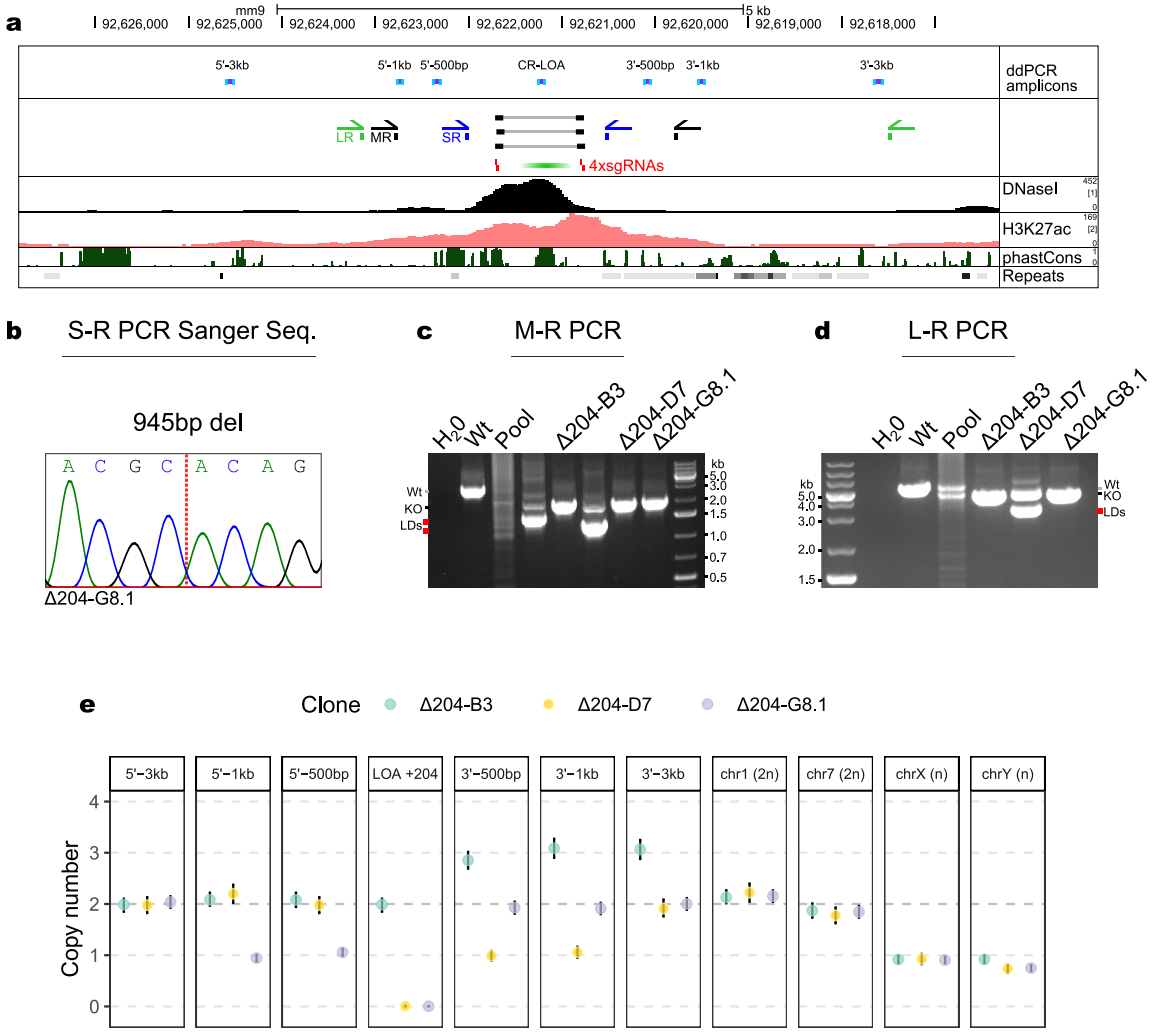


Figure 4.13 – Validation of *Runx1* +204 enhancer deletion clones. a) Locus map of targeting strategy used to delete *Runx1* enhancer +204. Locations of 4x sgRNAs used alongside Cas9^{D10A} nickaseare shown as red boxes, primers used in short-range (S-R), medium-range (M-R), and longer-range (L-R) PCR are indicated as half arrows. Digital droplet PCR (ddPCR) amplicons spaced over the enhancer region are shown as purple boxes with blue ends. Deletions mapped using S-R PCR and Sanger sequencing are indicated as grey boxes with black ends. The location of +204 enhancer as defined by Schütte et al. 2016 are indicated as green boxes. DNaseI-seq ([1] Vierstra et al. 2014) and H3K27ac ChIP-seq ([2] Schütte et al. 2016) shown were previously generated in 416B cells. PhastCons vertebrate conservation and repeats are indicated. b) S-R PCR Sanger sequencing trace from clone Δ110-E4.1 that was shown to harbour a 1565 bp deletion (a red dashed line indicates the position of the deletion). This clone contained only a single sequencing trace potentially indicating an allelic drop-out. c) M-R PCR gel image for Δ204-B3, Δ204-D7 and Δ204-G8.1 that were shown to harbour a knock-out for +204 enhancer by S-R PCR and a single Sanger sequencing trace. Wt and a grey bar indicates the size of the wild type allele, KO and a black bar indicates the size of the knock-out allele. *Legend continued on next page.*

Figure 4.13 – *Legend continued from previous page.* d) L-R PCR from clones Δ204-B3, Δ204-D7 and Δ204-G8.1 with identical labelling to c. e) ddPCR genotyping over the +204 enhancer region from Δ204-B3, Δ204-D7, and Δ204-G8.1. ddPCR droplets used to calculate relative copy numbers are shown in Appendix Figure 7.18. Relative copy numbers indicated were calculated relative to an internal control amplicon in each well, and to two different diploid chromosomes (chr1 and chr7) that were not targeted using $Cas9^{D10A}$ nickase. Amplicons targeting chrX and chrY were included as single copy controls. Three copies of the region 3' to *Runx1* +204 enhancer that were quantified in clone Δ204-B3 are indicated with blue stars above the copy number plots. Two copies of the +204 enhancer that were seen in clone Δ204-B3 is indicated with a red star above the copy number plot. Loss of *Runx1* +204 enhancer in clones Δ204-D7 and Δ204-G8.1 are indicated with a green star above the copy number plots. LDs that likely destroyed primer binding sites in clones Δ204-D7 and Δ204-G8.1 are indicated with orange stars above the copy number plots.

4.2.11 Redundant and non-redundant roles for *Runx1* enhancers during EHT *in vitro*

One knock-out clone for +110 enhancer (Δ110-E4.1) and two knock-out clones for +204 enhancer (Δ204-D7 and Δ204-G8.1) were used for directed haematopoietic differentiation *in vitro* according to our established *in vitro* EHT protocol (Figure 4.2). Alongside these enhancer-deleted clones, wild type 23C-RV cells were used as a positive control and 23C-RV cells engineered to lack a critical exon of Runx1 required for its DNA binding (Wang et al., 1996) were used as a negative control (23C-RV-Rd generated by Lucas Greder, Vincent Frontera, and myself). FACS quantification using haematopoietic surface markers CD41 and CD45 was used to investigate the blood forming potential of wild type 23C-RV cells, 23C-RV-Rd cells, and each of the enhancer knock-out clones (Figure 4.14 a). A two-way ANOVA test using genotypes (23C-RV, 23C-RV-Rd, Δ110, and Δ204) and cell populations (CD41- CD45-, CD41lo CD45-, CD41+ CD45-, CD41hi CD45-, CD41+ CD45+) as variables revealed that relative cell counts were significantly different between the genotypes ($F(1,11) = 7.67$, $p = 1.5\times10^{-8}$, $\eta^2 = 0.0898$). Specifically, post-hoc testing revealed that CD41-CD45- cells were increased in 23C-RV-Rd compared to wild type, reflecting a block in haematopoietic differentiation due to a lack of functional Runx1 protein (Figure 4.14 b lower bar graphs, *, Tukey's test, $p = 9.06\times10^{-9}$). CD41lo, CD41+, and CD41hi (all CD45-) populations were also observably but not significantly reduced in 23C-RV-Rd cells compared to 23C-RV wild type. None of the enhancer knock-out genotypes significantly differed from wild type in their ability to form the different haematopoietic cell populations (Figure 4.14 b lower bar graphs, *, Tukey's test, $p > 0.8$). Together this shows that deletion of a single *Runx1* enhancer (+110 or +204) does not significantly impact haematopoietic cell formation *in vitro*.

Next, using FACS we quantified *Runx1* expression indirectly using Runx1-Venus mean fluorescence intensity (MFI). Runx1-Venus was quantified separately in each of the discrete cell populations previously analysed and two way ANOVA was performed in order to test for differences in Runx1 expression levels associated with the different cell populations and genotypes examined. Overall, the main effect of genotype on Runx1-Venus MFI was significant ($F(1,3) = 14.2$, $p = 2.46\times10^{-7}$, $\eta^2 = 0.271$). Post-hoc testing using Tukey's test found that, compared to 23C-RV, Runx1-Venus expression levels were significantly reduced in Δ110 ($p = 3.45\times10^{-6}$) and 23C-RV-Rd ($p = 0.00793$) genotypes across all cell populations. The ANOVA test also found a significant interaction effect between genotype and cell population ($F(1,11) = 2.26$, $p = 0.0201$, $\eta^2 = 0.163$) indicating that genotype influenced Runx1-Venus MFI also within specific cell populations. Post-hoc testing revealed that Runx-Venus MFI was significantly reduced in Δ110 CD41lo CD45- emerging haematopoietic progenitor cells compared to 23C-RV wild type (Figure 4.14 b, top bar graphs, Tukey's test, *, $p < 0.02$). 23C-RV-Rd CD41+ CD45- cells also showed significantly reduced Runx1-Venus MFI compared to wild type 23C-RV (Figure 4.14 b, top bar graphs, Tukey's test, *, $p < 0.02$). Runx1-Venus MFI was not significantly different between Δ204 and 23C-RV ($p = 0.770$). Together, this FACS analysis suggested that Runx1 expression was reduced after +110 enhancer deletion, but not after +204 deletion.

Runx1 levels were also measured directly by immunofluorescence staining and confocal microscopy (Figure 4.14 c and d). All genotypes examined generated clusters of

CD31+ endothelial cells (Figure 4.14 c). 23C-RV-Rd cells showed no Runx1 staining and were useful as a negative control to quantify background levels of fluorescence (the Runx1 antibody used would normally bind to the exon deleted in 23C-RV-Rd cells). Blinded image quantification revealed that Runx1 protein levels were significantly reduced in single CD31+ CD41+ cluster cells of clone Δ110-E4.1 and 23C-RV-Rd cells (Figure 4.14 d, *, one way ANOVA with Tukey's post-hoc test $p < 0.001$). Haematopoietic colony formation assays (CFU-C) also revealed reduced colony forming potential in Δ110-E4.1 and 23C-RV-Rd cells (Figure 4.14 e, n = 1-3). Together, these data agree that Runx1 expression levels were reduced by +110 enhancer deletion and in Runx1-null cells, but not by +204 enhancer deletion.

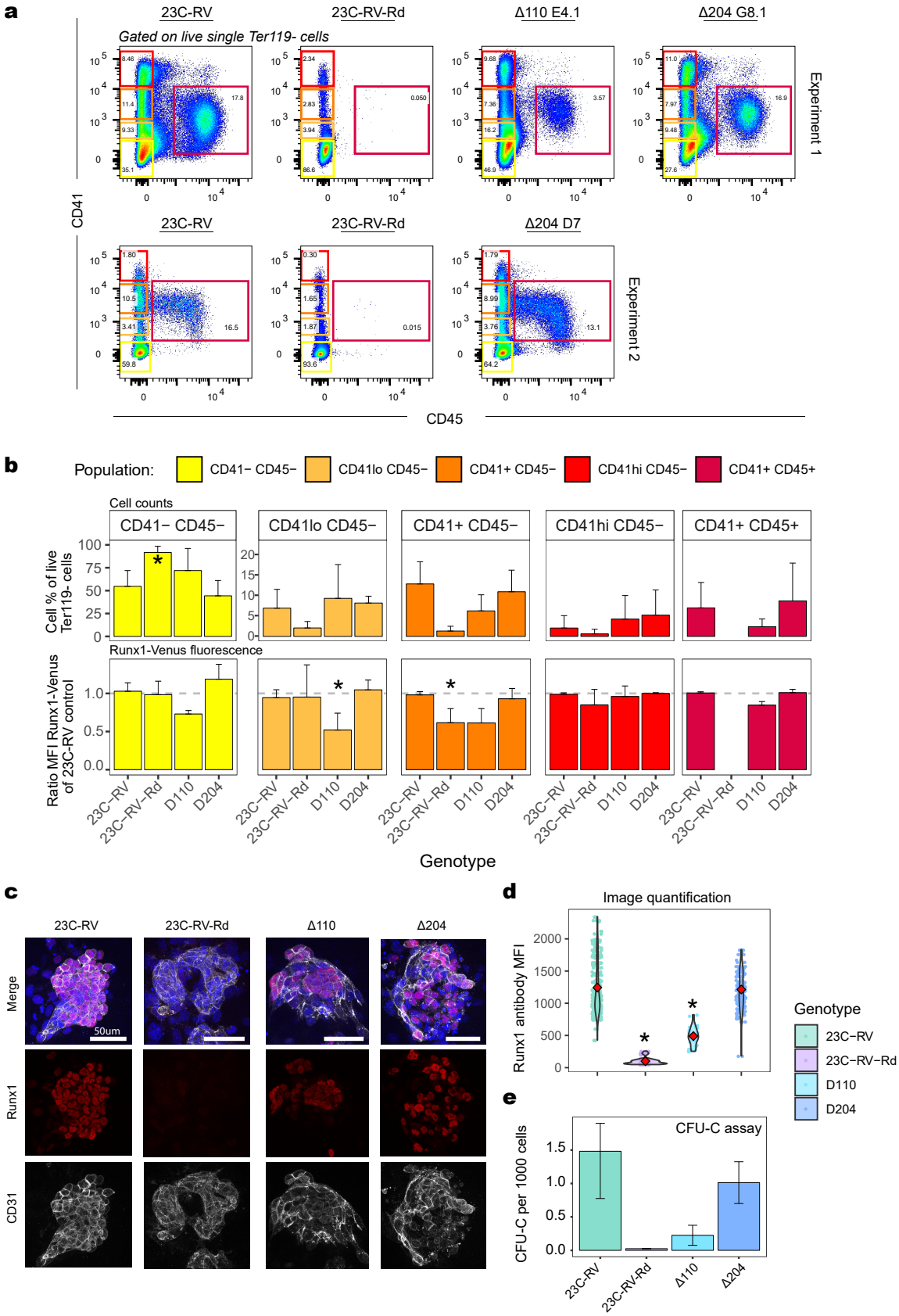


Figure 4.14 – Phenotypic characterisation of *Runx1* enhancer deletion clones. *Legend continued on next page.*

Figure 4.14 – *Legend continued from previous page.* a) FACS analysis of mESCs after 7 days of haematopoietic differentiation. Representative profiles from 23C-RV (wild type), 23C-RV-Rd (Runx1-null) Δ110-E4.1, Δ204-D7, and Δ204-G8.1 are shown from two independent experiments. Cells were gated on Ter119- live single cells. Differentiation and FACS analysis of cells was performed by Vincent Frontera and myself. FACS data analysis was performed by Vincent Frontera. b) Quantification of cell counts (as a percentage of Ter119- live single cells) and Runx1-Venus mean fluorescence intensity (MFI, as a fraction of 23C-RV control) in the different cellular populations indicated. 23C-RV n=5 independent experiments (first clone tested five times, second clone tested three times; values were averaged where both clones were tested on the same day); 23C-RV-Rd n=5 independent experiments (one clone tested five times); Δ110 n=3 independent experiments (one clone tested three times); Δ204 n=3 independent experiments (first clone tested once, second clone tested twice). Two-way ANOVA with Tukey's post-hoc test, *, $p < 0.02$. c) Representative immunofluorescence confocal images from one experiment showing staining of Runx1 and CD31 in haematopoietic clusters at day 7 of differentiation from each of the indicated genotypes. Immunostaining, confocal imaging, and image quantification was performed by Joe Harman, Christina Rode, and myself. d) Quantification of Runx1 antibody staining as mean fluorescence intensity (MFI) in single Runx1+ CD31+ CD41+ cells. Each dot represents an individual cell that was quantified, the violin plot represents the distribution and the red dot represents the mean of all the cells. Genotypes are indicated by a different colour plot according to the key. One-way ANOVA was performed ($F(1,3)=447.3$, $p < 0.001$) and post-hoc testing was done with the Tukey's test (*, $p < 0.05$ compared to 23C-RV). 23C-RV: n = 3 independent experiments (one clone tested three times), 109 cells quantified. 23C-RV-Rd: n = 3 independent experiments (one clone tested three times), 45 cells quantified. Δ110: n = 1 independent experiment (one clone tested once), 35 cells quantified. Δ204; n = 2 independent experiments, (one clone tested twice), 82 cells quantified. e) Colony formation assays were done for each of the genotypes analysed and colonies counted after 10 days. 23C-RV; n = 4 independent experiments (first clone tested three times, second clone tested once). 23C-RV-Rd; n = 1 independent experiment (one clone tested once). Δ110; n = 2 independent experiments (one clone tested twice). Δ204; n = 2 independent experiments (two clones each tested once) Error bars represent the minimum and maximum values.

4.3 Chapter conclusions and discussion

In this chapter, I defined the dynamic activities of *Runx1* enhancers during EHT *in vitro*. Using an *in vitro* directed differentiation protocol, we generated cellular intermediates of EHT and performed ATAC-seq on them. Previous studies utilised immortalised cell lines to study chromatin marks at *Runx1* enhancers (Nottingham et al., 2007; Ng et al., 2010; Schütte et al., 2016). However, my analysis of our ATAC-seq and other publicly available chromatin marks in HE and HP cells demonstrated that the *Runx1* enhancers have tissue-specific chromatin accessibility and TF binding during EHT *in vitro*. Similar differentiation protocols have previously been used to isolate and study distinct cellular intermediates during endothelial-to-haematopoietic transition (EHT) (Lancrin et al., 2009; Sroczynska et al., 2009b; Lancrin et al., 2012; Pearson et al., 2015; Goode et al., 2016).

In vitro differentiation systems are advantageous over *in vivo* systems due to allowing large cell numbers that are required for most chromatin assays like ChIP-seq to be generated relatively painlessly. Moreover, these systems are useful for allowing rapid screening of genetic perturbations. *In vitro* differentiation systems are not without their disadvantages, however. One issue is that they are generally considered to only recapitulate the yolk-sac definitive (second) wave of embryonic haematopoietic development (Ditadi et al. 2017, Figure 1.2). *In vitro* haematopoietic mESC differentiation was also less synchronised than a developing embryo, complicating the analysis. Both variables can lead to large differences in the 'quality' of the differentiation. Specifically, this led to several failed experiments when attempting to generate samples for ATAC-seq which limited the total number of biological samples that were obtained. This likely contributed to the relatively poor signal-to-noise ratio of our ATAC-seq data. Reproducibility of differentiations may be increased by using 2i medium to maintain mESCs (Tamm et al., 2013).

Technical factors to do with the transposition reaction itself may also have reduced the signal-to-noise ratio of our ATAC-seq data. Firstly, the ATAC-seq protocol involved several centrifugation steps cells after FACS sorting which could have triggered cell death. This may be prevented by sorting cells directly into transposition buffer (O'Byrne et al., 2019). Second, to preserve cells, the exact number of cells used for ATAC-seq was estimated from the sorted event counts of the sorter. It is known that DNA fragment sizes depend on the ratio of cells to transposase enzyme (Buenrostro et al., 2015), and the true cell number may have been different from the counter on the sorter and not have been optimal for the amount of transposase used. Counting cells using a haemocytometer could improve this (though cell death may then still be an issue). Third, the transposition reaction may have been carried out for too long. EDTA was added to quench the enzyme, but this step still took some time, which may have increased the background signal in the samples. Using an automated pipette might allow faster quenching of the enzyme. In combination, these measures may improve the signal-to-noise ratio of highly purified cell populations in future experiments.

Despite a relatively low signal-to-noise ratio, it was clear that during *in vitro* EHT the patterns of endogenous enhancer activity generally agreed with the previously identified patterns of activity determined in enhancer-reporter assays during EHT *in*

vivo (Nottingham et al., 2007; Bee et al., 2010; Schütte et al., 2016). This suggests that the activities of haematopoietic *Runx1* enhancers in enhancer-reporter studies may accurately recapitulate the endogenous enhancer activities. However, some technical limitations should be considered when interpreting the observed increase in +23 enhancer accessibility. 23C-RV cells contain the +23 enhancer knocked-in to the *Col1a1* locus (Lucas Greder, unpublished). Reads originating from the two copies of the +23 enhancer in the endogenous *Runx1* locus or the two copies in the +23-Cherry transgene cannot be distinguished and both will be mapped to the endogenous locus. Previously generated chromatin accessibility data in HE cells derived from mESCs lacking the +23 transgene (Goode et al., 2016) showed accessibility at the endogenous +23 enhancer in the *Runx1* locus, suggesting that the endogenous enhancer was active in 23C-RV cells also. However, since the enrichment of reads at +23 was higher in 23C-RV-derived HE cells compared to HE cells lacking the +23 transgene this suggests that a proportion of the reads at the endogenous +23 enhancer in 23C-RV-derived HE cells could be attributed to the transgene. One possible solution to this could be to generate a mutated +23 enhancer with silent point mutations at least every 30 bp. This could facilitate non-spurious mapping of reads to a custom genome containing the transgenic and endogenous enhancers since ATAC-seq samples are usually sequenced on a 35+35 bp paired-end Illumina® sequencing run. Mutations would need to be carefully selected not to impact important regulatory upstream TFBS, however, and the impact of the mutations to the +23 enhancer would need to be validated experimentally. Alternatively, the +23 enhancer from another species could be used (with or without additional sequence mutations modification) if it was sufficiently different in sequence to the mouse enhancer. Enhancers are functional across species in this way—one prior study replaced a mouse enhancer of *Shh* with that of a snake and was able to recapitulate snake-like limb *Shh* expression and phenotype (Kvon et al., 2016)—but due to the high levels of evolutionary sequence conservation of the +23 enhancer finding an enhancer that is distinguishable may be challenging.

Many upstream TFs that bound *Runx1* enhancers during EHT *in vitro* have previously implicated in *Runx1* transcriptional regulation including Ets factors (Nottingham et al., 2007; Schütte et al., 2016), Gata2 (Robert-Moreno et al., 2005; Burns et al., 2005; Nottingham et al., 2007), and Scl (Nottingham et al., 2007; Pimanda et al., 2007; Schütte et al., 2016). This reflects a core set of heptad haematopoietic TFs binding at these enhancers which have previously been shown to drive many haematopoietic genes in immortalised HPC-7 cells (Wilson et al., 2010a). However, HE and HP samples used for ChIP-seq were mixed populations containing millions of cells and TF binding at each enhancer may have only occurred in a subset of each population. To determine the fraction of cells in each population in which a TF was bound single-cell ChIP-seq (Rotem et al., 2015) or single cell CUT&RUN (Hainer et al., 2019) could be used. However, absence of evidence of TF binding is not equivalent to evidence of absence of TF binding; care would still need to be taken in order to interpret any single cell TF binding data. I also turned to meta digital DNaseI footprinting in HE cells coupled with consensus TFBS and evolutionary conservation analysis to identify novel upstream regulators of *Runx1* during EHT. Sometimes, DNaseI footprints were identified in regions without sequence conservation and no

obvious consensus TFBS. One explanation could be noise in the DNaseI footprinting data due to DNaseI cutting biases (Schwessinger et al., 2017). Alternatively, these could represent species-specific motifs, or a real TFBS that is not important/bound in that enhancer and so is not conserved. It would be interesting to mutate some of the DNaseI footprints that were not conserved to see if they have a role to play in enhancer function. Additionally, some conserved consensus binding motifs were not identified as DNaseI footprints in HE and 416B cells. A reasonable explanation could be that these motifs would be bound by TFs (and would generate footprints) in different cell types.

Deeply conserved Klf, Cebp, and AP-1 motifs were shown to be unique to the +23, +110, and +204 enhancers, respectively. Luciferase assays suggested that each of these motifs were important for enhancer function, suggesting that these factors may regulate *Runx1* transcription. However, the conclusions that can be drawn from the luciferase experiments alone are somewhat limited. Luciferase assays are highly artificial, and the constructs were expressed in 416B cells which do not express the same set of upstream TFs as would be present in cells undergoing EHT. However, expression data showed that upstream TFs capable of binding these motifs were expressed in cellular stages when the enhancer-reporters were active, making it seem plausible that they do regulate *Runx1* transcription. Moreover, there is a precedent for many of these factors being involved in regulating similar haematopoietic genes. For example, *Klf4* was upregulated in HE cells and has previously been implicated in regulating *Vegfa*, Notch signaling and angiogenesis (Hale et al., 2014; Wang et al., 2015; Li et al., 2019; Boriushkin et al., 2019). It has also been shown to play a role in endothelial-to-mesenchymal transition (EMT) (Tiwari et al., 2013), a cellular process sharing many similarities with EHT (Ottersbach, 2019). Interestingly, *Runx2* was shown to be regulated by Klf10 in osteoblasts (Hawse et al., 2011). Sp1, which binds the same motifs as Klf family TFs (Cao et al., 2010) was shown to repress *Runx1* during EHT (although Sp1 binding at *Runx1* was not directly examined Gilmour et al. 2014). Further studies should explore a role for Klf/Sp1 family genes which may be direct upstream transcriptional regulators of *Runx1* during EHT, plausibly via the +23 enhancer.

Another TF that is worth exploring as an upstream regulator of *Runx1* is AP-1. Since AP-1 binding sites are redundant (Fonseca et al., 2019) and the complex contains many different interchangeable subunits (Shaulian and Karin, 2002), it is unclear exactly which members may be directly important for regulating *Runx1* regulation. In support of a role for AP-1 upstream of *Runx1*, several AP-1 family members were upregulated during EHT both *in vitro* and *in vivo*. Moreover, Junb was previously shown to be upstream of *runx1* during haematopoietic development in zebrafish (although direct binding was not assessed, Li et al. 2015). Interestingly, endogenous *Runx1* expression was achieved during the conversion of adult endothelium to HSCs *in vitro* by the transient expression of AP-1 family member *Fosb* alongside *Gfi1*, *Runx1*, and *Spi1* (Lis et al., 2017). Together, AP-1 genes seem likely upstream regulators of *Runx1* transcription, possibly mediated by binding to the +204 enhancer.

The last new potential upstream regulator of *Runx1* identified were Cebp family genes. A deeply conserved Cebp motif was shown to be important for *Runx1* +110 enhancer

function, and Cebpβ bound to the enhancer in mESC-derived HP cells (Goode et al., 2016). *RUNX1* was shown to be upstream of *CEBP*α before, (Guo et al., 2012), but to the best of my knowledge Cebp genes have not been suggested to be upstream of *Runx1*. Since all Cebp factors recognise the same motif, it is possible that multiple Cebp factors regulate *Runx1*. Establishing which upstream factors are responsible for transcriptional regulation of *Runx1* could inform *in vitro* differentiation protocols seeking to generate HSCs *de novo*.

Analysis of chromatin marks and TF binding during EHT *in vitro* confirmed that the haematopoietic enhancers identified in transgenics studies (Nottingham et al., 2007; Bee et al., 2010; Swiers et al., 2013a; Schütte et al., 2016) were dynamically regulated during EHT. This led us to examine their functional requirements during EHT *in vitro*. Unexpected LDs frequently occurred while using CRISPR/Cas9. Sequencing many of these breakpoints led to the discovery that microhomologies were prevalent at Cas9-induced LDs and I showed that they cannot be used to predict LDs (Owens et al., 2019). This finding is potentially significant for the field for several reasons. Firstly, recent attempts to predict shorter Cas9 deletions (< 50 bp) have relied heavily on microhomologies (Allen et al., 2018; Shen et al., 2018). Instead of microhomologies being good predictors of LDs, distance from cut sites was a reasonable predictor of LD frequency in populations of cells (Owens et al., 2019). This knowledge will be useful for designing gene-editing and screening approaches for research or therapeutic applications in the future. Second, it was recently shown that pharmacologically inhibiting MMEJ reduced the frequency of recovered alleles containing microhomologies at their junctions (Iyer et al., 2019). It remains to be seen whether transient MMEJ inhibition could be used to reduce the occurrence of LDs when gene editing for research or therapeutic applications.

Due to finding LDs in clones that appeared homozygous knock-out for +23, it was not possible to determine the functional role of this enhancer in this way. One study saw reduced RUNX1 protein levels after targeting the *RUNX1* +23 enhancer in a human AML cell line (Mill et al., 2019). However, the authors targeted a population of cells with nine sgRNAs at once and did not perform genotyping of the resulting cell population (Mill et al., 2019). It has previously been shown that multiple adjacent DSBs might increase LD frequency (Owens et al., 2019), making the effect on RUNX1 protein levels impossible to attribute entirely to enhancer-mediated transcriptional effects. Cells may also have reduced protein levels after suffering loss of *RUNX1* exons or promoters. Moreover, after growing for several days, clonal selection may have taken place leading to a skew in the resulting RUNX1 levels. Despite these limitations, previous work in our laboratory saw a reduction in haematopoietic output (measured by CFU assays) in mESC lines where one copy of the +23 enhancer had been deleted by HR. This suggests that the +23 enhancer is functionally required for haematopoietic development *in vitro* and, therefore, likely contributes to *Runx1* transcription levels during EHT.

Differentiation of thoroughly genotyped homozygous enhancer knock-out clones revealed that loss of the +110 enhancer decreased Runx1-Venus and Runx1 protein levels $\sim$30-50% in cells undergoing EHT. It is important to note, however, that these results are based on a single clone. Further clones are being examined to ensure

that the reduced *Runx1* levels observed are not due to non-specific deficits in this particular mESC line. Moreover, *Runx1* transcripts should be measured directly in wild type and enhancer knock-out cells using RNA-seq or qRT-PCR. Previously, +23 and +110 enhancers were suggested to form a SE (Mill et al., 2019; Gunnell et al., 2016; Schuijers et al., 2018; Hnisz et al., 2017). The fact that haematopoietic cells were generated (albeit at a slightly lower frequency by CFU analysis) and Runx1 levels were only modestly impacted by +110 enhancer deletion argues against this. Therefore, *Runx1* enhancers are more likely to be acting in an additive manner, as was previously observed for enhancers of α-globin (Hay et al., 2016) and β-globin (Bender et al., 2012).

Interestingly, deletion of the +204 enhancer appeared to have no effect on haematopoietic differentiation and Runx1 levels. This might imply that Runx1 enhancers have distinct functional requirements during EHT and that the +204 is not required in this process. Differential enhancer activities assessed by chromatin marks and TF binding in the endogenous locus during EHT agrees that the enhancers are not exact duplicates of each other. Since the +204 enhancer knock-out clones showed no phenotype, it is possible that another enhancer—such as the +23 enhancer which is also active in HE—compensated for its loss and that redundancy exists between the enhancers.

5. Regulation of *Runx1* alternative promoter choice

5.1 Introduction

Runx1 promoter usage is a tightly regulated process during mouse and human development (Miyoshi et al., 1995; Ghozi et al., 1996; Telfer and Rothenberg, 2001; Pozner et al., 2007; Sroczynska et al., 2009a; Bee et al., 2009b, 2010). While regulators of this key developmental process are poorly understood, some evidence implicates CTCF/cohesin. Firstly, both *Runx1* promoters and intronic regions bound CTCF and rad21 in zebrafish (Marsman et al., 2014) and this was also the case in human Jurkat cells (Schuijers et al., 2018). Secondly, *runx1* promoter usage was altered after whole-organism morpholino-mediated CTCF or rad21 depletion during haematopoietic development in zebrafish (Marsman et al., 2014). *Runx1* was also upregulated after four days of CTCF depletion in mESCs was (Nora et al., 2017). However, the observed defects in *Runx1* transcription after CTCF/Rad21 knock-down (Marsman et al., 2014; Nora et al., 2017) could be mediated via a direct effect on *Runx1* or an indirect effect via an upstream regulator of *Runx1*. In addition, the cellular identity of the cells may be altered by genome-wide loss of CTCF/cohesin (such as inducing differentiation in CTCF depleted mESCs), altering expression levels of *Runx1* at the population level. Therefore, a locus-specific analysis of the role of specific CTCF binding sites on *Runx1* transcription is lacking.

To investigate a possible role for CTCF in regulating *Runx1* promoter usage, tissue-specific and constitutive CTCF and Rad21 binding was examined at the *Runx1* locus in three different cell lines, E14, 416B, and HPC7. *In vitro* insulator assays were done to investigate potential differences in the insulator properties of tissue-specific and constitutive CTCF binding sites. RNA sequencing was used to investigate the relationship between CTCF binding activity and promoter activity in the three cell lines. The relationship between DNA methylation, CTCF binding, and promoter activity was investigated using bisulfite sequencing. Correlations between CTCF binding and promoter activity were then also explored genome-wide.

5.2 Results

5.2.1 Binding of CTCF is dynamic at the *Runx1* locus

To identify sites that might be involved in regulating *Runx1* transcription and promoter usage during developmental haematopoiesis, CTCF and cohesin binding was examined at the *Runx1* locus in undifferentiated E14 mESCs and two different haematopoietic progenitor cell types (416B myeloid progenitor cells Dexter et al. 1979, and HPC7 haematopoietic progenitor cells (Pinto do O et al., 1998b)). As was noted in Chapter 3 Figure 3.2, CTCF binds at multiple sites within the *Runx1* domain including at the boundaries of the regulatory domain (Figure 5.1), and at the *Runx1* promoters

(Figure 5.1). CTCF enrichment was similar at many sites in all three cell types (blue circles, Figure 5.1 b and c) including at the centromeric and telomeric boundary CTCF sites. Rad21 binding was seen at the boundary CTCF sites in HPC7 cells but read depth was too low to assess Rad21 enrichment in E14 and 416B cells. Interestingly, some sites were bound by CTCF in a tissue-specific manner (black open circles, Figure 5.1 b and c) including the sites at both *Runx1* promoters (red stars, Figure 5.1 c). The P2 promoter bound CTCF and Rad21 in E14 and 416B cells, but not in HPC7 cells (Figure 5.1 c). In contrast, P1 was bound by CTCF in 416B and HPC7 cells, but not in E14 mESCs. Rad21 bound to P1 in HPC7 cells while no significant enrichment was seen in 416B or E14 mESCs. Together, these data show that in the cell types examined CTCF bound constitutively at sites towards the boundaries of the *Runx1* domain, while binding was more dynamic at the *Runx1* promoters.

5.2.2 *Runx1* CTCF sites act as insulators *in vitro*

One possible role for CTCF binding sites is that they could act as insulators (Chung et al., 1993; Bell and Felsenfeld, 1999; Zhao and Dean, 2004) and thus affect promoter-enhancer interactions in the wider locus. To begin to address this, we asked whether there was a difference in CTCF-dependent insulator properties of constitutive and dynamic CTCF binding sites in the Runx1 locus. This was examined using episomal *in vitro* insulator assays (Chung et al., 1993; Bell and Felsenfeld, 1999). One constitutively bound site was chosen (-468-CTCF at the telomeric boundary of the *Runx1* domain), and one dynamic CTCF site (P2-CTCF upstream of the P2 promoter). Each was cloned into a series of luciferase enhancer-reporter plasmids, or with the chicken HS4 insulator element used as a positive control (Zhao and Dean, 2004), and electroporated into 416B cells (Figure 5.2 a).

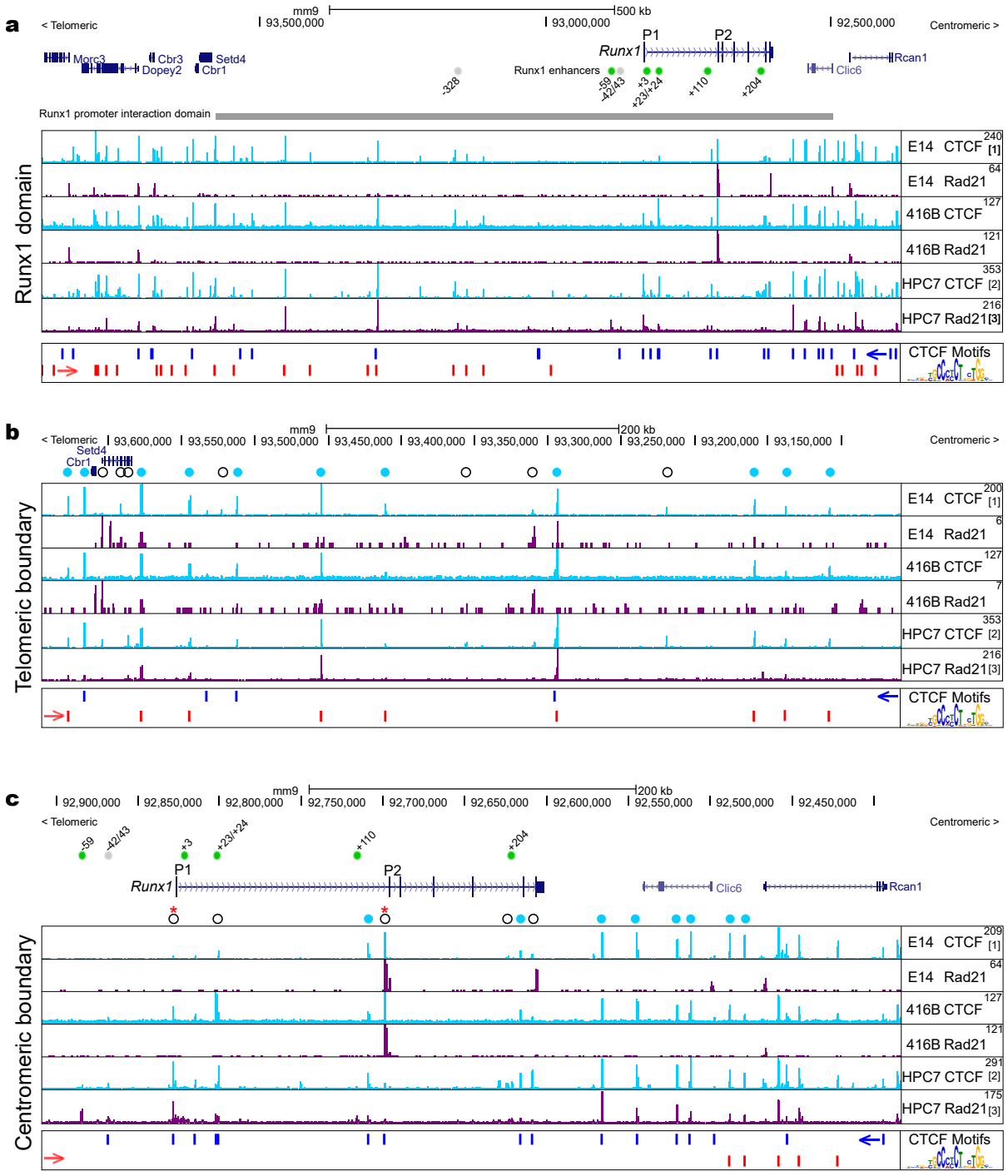


Figure 5.1 – Dynamic and constitutive CTCF binding in the *Runx1* locus. *Legend continued on next page.*

Figure 5.1 – *Legend continued from previous page.* a) Annotation showing *Runx1* gene, P1 and P2 promoters, and enhancers previously shown to drive reporter expression in haematopoietically active sites (green circles) or putative enhancers that did not drive expression to haematopoietically active sites (grey circles) in embryos (Nottingham et al., 2007; Schütte et al., 2016). The approximate extent of the *Runx1* promoters interaction domain defined by Capture-C is indicated by a grey box. CTCF and Rad21 binding is shown in E14 mESCs ([1], Handoko et al. 2011), 416B cells, and HPC7 cells ([2],Calero-Nieto et al. 2014; [3], Wilson et al. 2016) over the entire *Runx1* regulatory domain. The orientation of *de novo* annotated CTCF motifs are shown in blue and red. b) Close up view of the telomeric boundary region showing CTCF and Rad21 binding in E14, 416B and HPC7 cells. Blue circles above the tracks indicate peaks with similar binding in all three cell types, black outlined circles indicate CTCF sites with different binding patterns between the cell types. c) Close up view of the *Runx1* gene, showing its enhancers, promoters, and the centromeric boundary with *Clic6/Rcan1*. Blue circles above the tracks indicate peaks with similar binding in all three cell types, black outlined circles indicate CTCF sites with different binding patterns between the cell types. CTCF binding sites upstream of the promoters are indicated by red stars. UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_CTCF_wide`.

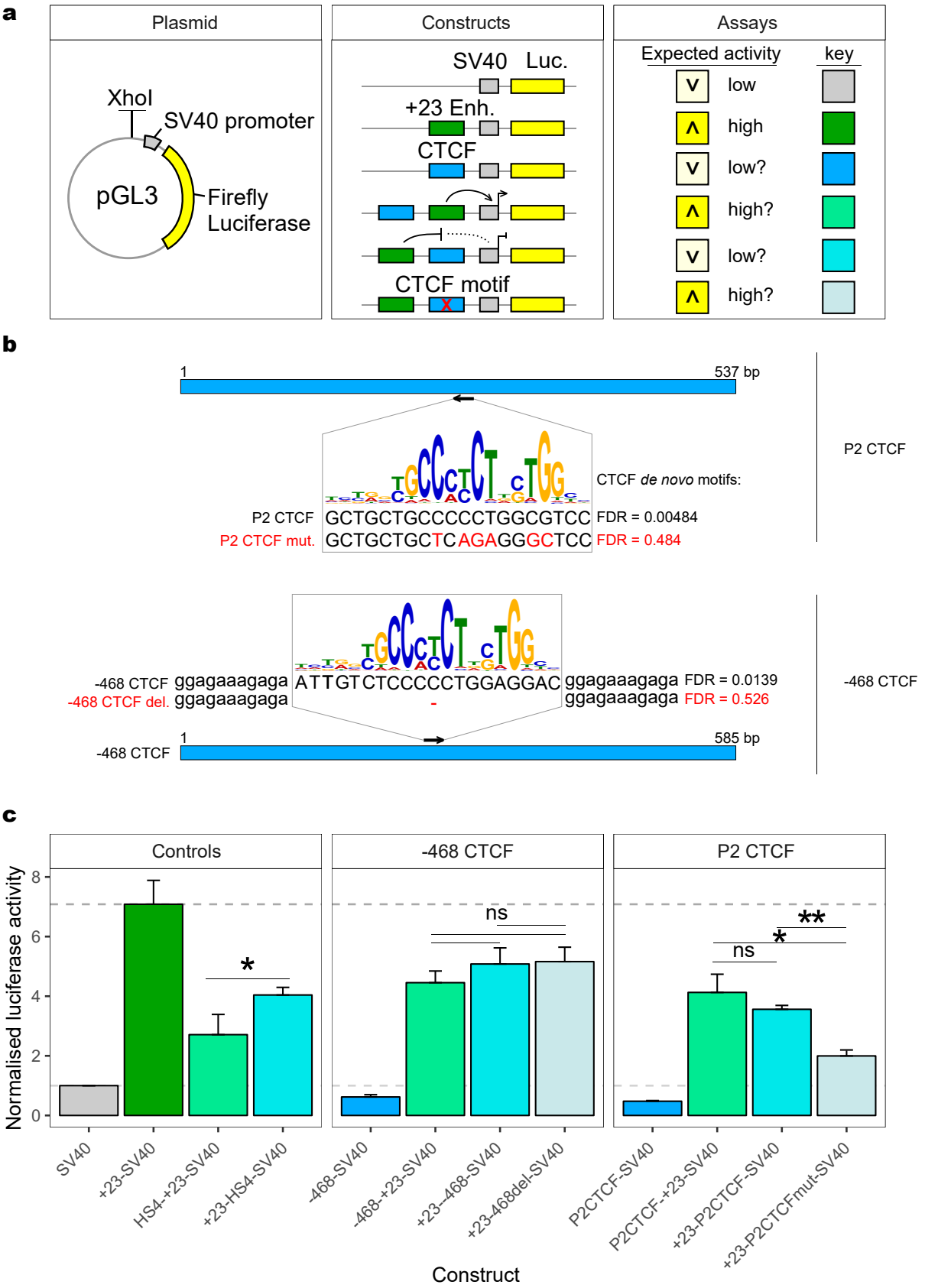


Figure 5.2 – *In vitro* insulator assays to examine insulator functions of CTCF sites in the *Runx1* domain. *Legend continued on next page.*

Figure 5.2 – *Legend continued from previous page.* a) Schematic showing pGL3 luciferase plasmids containing *Runx1* +23 enhancer alongside different CTCF binding sites in different combinations with or without intact predicted CTCF binding sites. The expected activity of the constructs based on prior studies (Chung et al., 1993; Bell and Felsenfeld, 1999) is shown and the colour key corresponds to the bars in c. b) Map of the P2-CTCF and -468 CTCF sites that were cloned into insulator constructs. The orientation of predicted CTCF binding sites that matched a *de novo* CTCF motif identified in 416B CTCF ChIP-seq are indicated by black arrows (FDR > 0.02). c) Results of luciferase assays to reveal insulating properties of CTCF binding sites and the chicken HS4 positive control insulator (Zhao and Dean, 2004). One-way ANOVA showed that each of the constructs showed significantly reduced luciferase activity compared to +23 enhancer alone ($F(1,11)=66.3$, $p = 3.29\times10^{-15}$, $\eta^2 = 0.968$, Tukey's post-hoc test, $p < 7.2\times10^{-4}$). Individual post-hoc tests were performed to test between different plasmids (ns, $p > 0.7$, *, $p < 0.05$, **, $p = 0.0082437$). n=3 independent experiments per construct with three technical replicates per experiment. Cloning of constructs was performed by Akin Bucakci and myself and Luciferase assays were performed by Akin Bucakci.

Based on previous work (Bell and Felsenfeld, 1999), CTCF binding sites placed between the enhancer and promoter were expected to have the largest insulating effect, while a reduced insulating effect was expected when CTCF sites were inserted upstream of the enhancer (Figure 5.2 a, Bell and Felsenfeld 1999). A single predicted CTCF binding motif was identified within each of the ∼550 bp sequences by *de novo* motif calling algorithm (FDR < 0.02) and mutation or deletion of the predicted CTCF motif abolished the predicted binding sites (FDR > 0.4, Figure 5.2 b). After normalising to a control plasmid (containing renilla luciferase to account for electroporation efficiency), each of the CTCF sites alongside the *Runx1* +23 enhancer exhibited an insulating effect. Each of the insulator constructs showed reduced luciferase expression compared to +23 enhancer alone (Figure 5.2 c, one-way ANOVA, $F(1,11)=66.3$, $p = 3.29\times10^{-15}$, $\eta^2 = 0.968$, Tukey's post-hoc test, $p < 7.2\times10^{-4}$). Unexpectedly, when CTCF sites were located upstream of the enhancer (and not between the enhancer and SV40 promoter) the insulating effect was still seen (Figure 5.2 c). Moreover, the HS4 insulator element showed a greater insulating effect when placed upstream of the +23 enhancer compared to when placed between the enhancer and promoter (Figure 5.2 c, Tukey's post-hoc test, *, $p < 0.05$). Placing the P2-CTCF or -468 CTCF upstream or downstream of the +23 enhancer produced the same insulating effect (Figure 5.2 c, Tukey's post-hoc test, ns, $p > 0.7$). Deleting the predicted CTCF motif made no difference to the insulating effect of the -468 CTCF site (Figure 5.2 c, Tukey's post-hoc test, ns, $p > 0.7$), while mutating the P2-CTCF site increased the insulating effect of the element and reduced luciferase expression (Figure 5.2 c, Tukey's post-hoc test, **, $p = 0.0082437$). Observably, -468 and P2-CTCF sites on their own reduced baseline expression from the SV40 promoter alone, although this was not significant (Figure 5.2 c, Tukey's post-hoc test, $p > 0.9$). Together, both constitutive and dynamic CTCF binding sites elicited insulator properties *in vitro* that did not depend on CTCF site location or predicted CTCF binding sites.

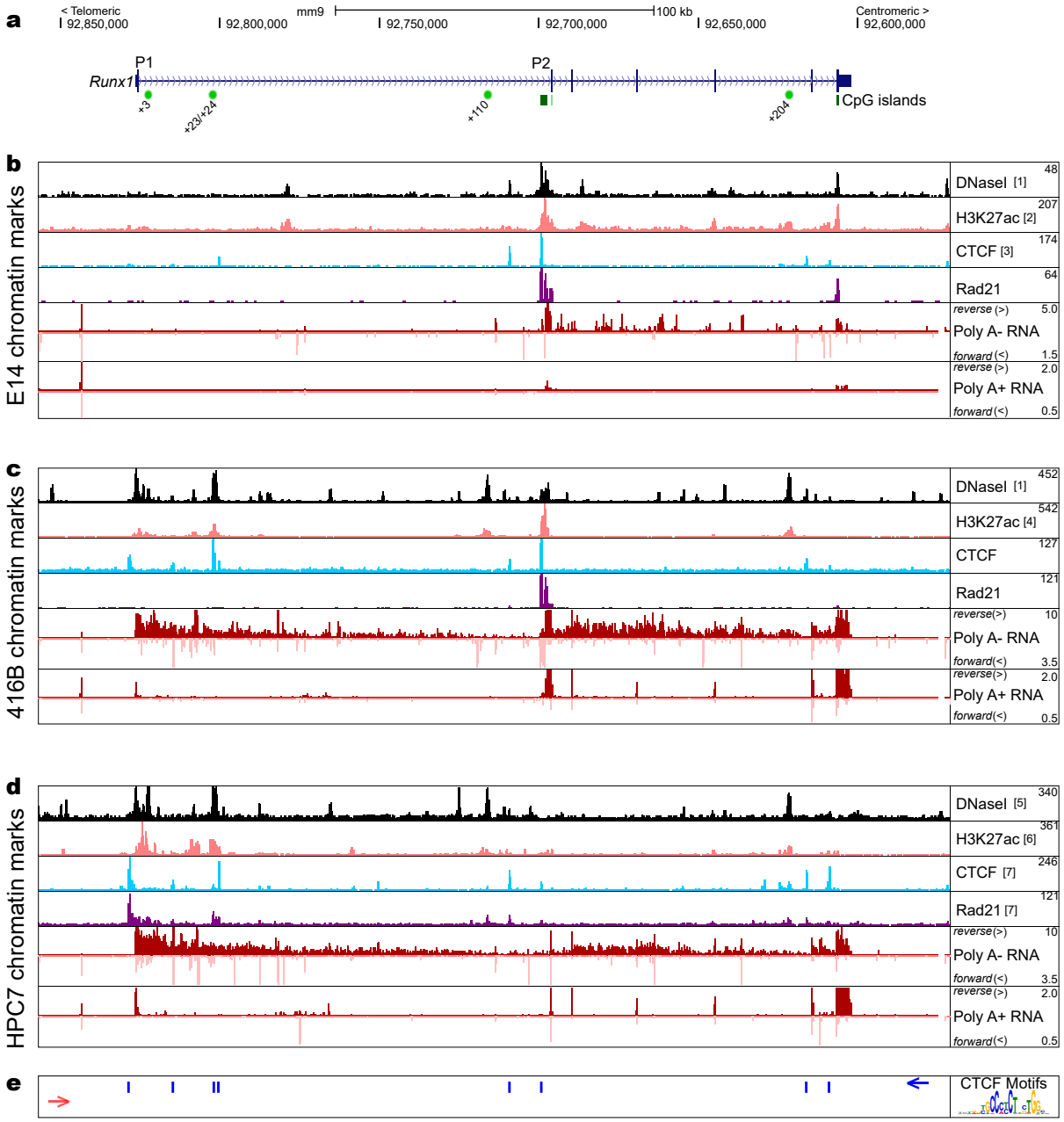


Figure 5.3 – CTCF binding at the *Runx1* promoters is correlated with promoter activity. a) Annotation showing *Runx1* gene, P1 and P2 promoters, and enhancers previously shown to drive reporter expression in haematopoietic sites in embryos (green circles) (Nottingham et al., 2007; Schütte et al., 2016). CpG island annotation from UCSC is shown. b) Chromatin marks in undifferentiated E14 mESCs including DNaseI ([1], Vierstra et al. 2014) H3K27ac ([2], Wamstad et al. 2012), CTCF ([3], Handoko et al. 2011), Rad21, poly A minus and plus RNA-seq. c) Chromatin marks in 416B myeloid progenitor cells including DNaseI ([1], Vierstra et al. 2014) H3K27ac ([4], Schütte et al. 2016), CTCF, Rad21, poly A minus and plus RNA-seq. d) Chromatin marks in HPC7 haematopoietic progenitor cells including DNaseI ([5], Wilson et al. 2010a) H3K27ac ([6], Calero-Nieto et al. 2014), CTCF ([6], Calero-Nieto et al. 2014), Rad21 ([7], Wilson et al. 2016), poly A minus, and plus RNA-seq. e) The orientation of *de novo* annotated CTCF motifs are shown in blue. UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_Methylation_wide`

5.2.3 CTCF binding is correlated with promoter activity and DNA methylation at *Runx1* promoters

Next, we asked whether Runx1 alternative promoter activity might be associated with differential CTCF binding at the promoters. In E14 and 416B cells, a distinct CTCF peak can be observed ∼1.5 kb upstream of the P2 promoter (Figures 5.3 a-c, 5.4). *De novo* CTCF motif analysis revealed that the motif was oriented in the opposite direction to transcription (Figures 5.3, 5.4). Rad21 also bound at this CTCF peak and across the P2 promoter region (Figure 5.4 b and c). Active chromatin marks–DNaseI-seq, and H3K27ac–suggested that the P2 was active or poised in both E14 and 416B cells (Figures 5.3 a-c, 5.4). Indeed, poly A+ and poly A- RNA-seq showed that transcription was occurring from the P2 in both cell types (Figure 5.4 b and c). Interestingly, poly A- RNA-seq showed bidirectional transcription emanating from the P2 TSS, including anti-sense transcription over the P2 CpG island (CGI) (Figures 5.3 a-c, 5.4). In contrast, in HPC7 cells the P2 promoter was inactive. RNA-seq reads were primarily only seen over exon 3 in these cells, and no bidirectional transcription could be seen from the P2 promoter (Figures 5.3 a-c, 5.4). Moreover, DNaseI-seq, and H3K27ac, CTCF, Rad21 enrichment was much lower than in the other cell types (Figures 5.3 d, 5.4). Targeted bisulfite sequencing showed that the P2 promoter CGI was demethylated in E14 and 416B cells and was methylated in HPC7 cells (Figures 5.3, 5.5). Interestingly, the CTCF motif itself was also demethylated in E14 and 416B cells and methylated in HPC7 cells.

The Runx1 P1 promoter was also flanked by a CTCF peak ∼1.8 kb upstream of its TSS in 416B and HPC7 cells, but not in E14 cells. Again, the motif was oriented in the opposite direction to transcription (Figures 5.3, 5.5). In 416B and HPC7, but not in E14, the promoter was active as shown by DNaseI hypersensitivity and H3K27ac enrichment (Figures 5.3 a-d, 5.5 a-d). Interestingly, anti-sense transcription (to the orientation of *Runx1* transcription) was observed in the first intron of *Runx1* in 416B and HPC7 cells and was most prominent in the region of the +23 and +3 enhancers and might represent non-coding enhancer RNAs (Figures 5.3 c and d, 5.5 c and d, Kim et al. 2010; Hah et al. 2011; Wang et al. 2011; Kowalczyk et al. 2012). Another explanation for the observed anti-sense transcripts could also be imperfect strandedness in the RNA-seq libraries. However, this seems unlikely because at the P2 promoter, bidirectional transcription on opposite strands could clearly be resolved at the P2 TSS (Figure 5.4, b and c). Targeted bisulfite sequencing revealed that in 416B and HPC7 cells the P1 promoter was demethylated (Figures 5.3 c and d, 5.5 c and d). In contrast, in E14 mESCs, the inactive P1 promoter region was methylated (Figures 5.3 b, 5.5 b). Therefore, CTCF only bound upstream of the *Runx1* promoters when they were active and demethylated, suggesting a relationship between CTCF binding, methylation, and promoter activity.

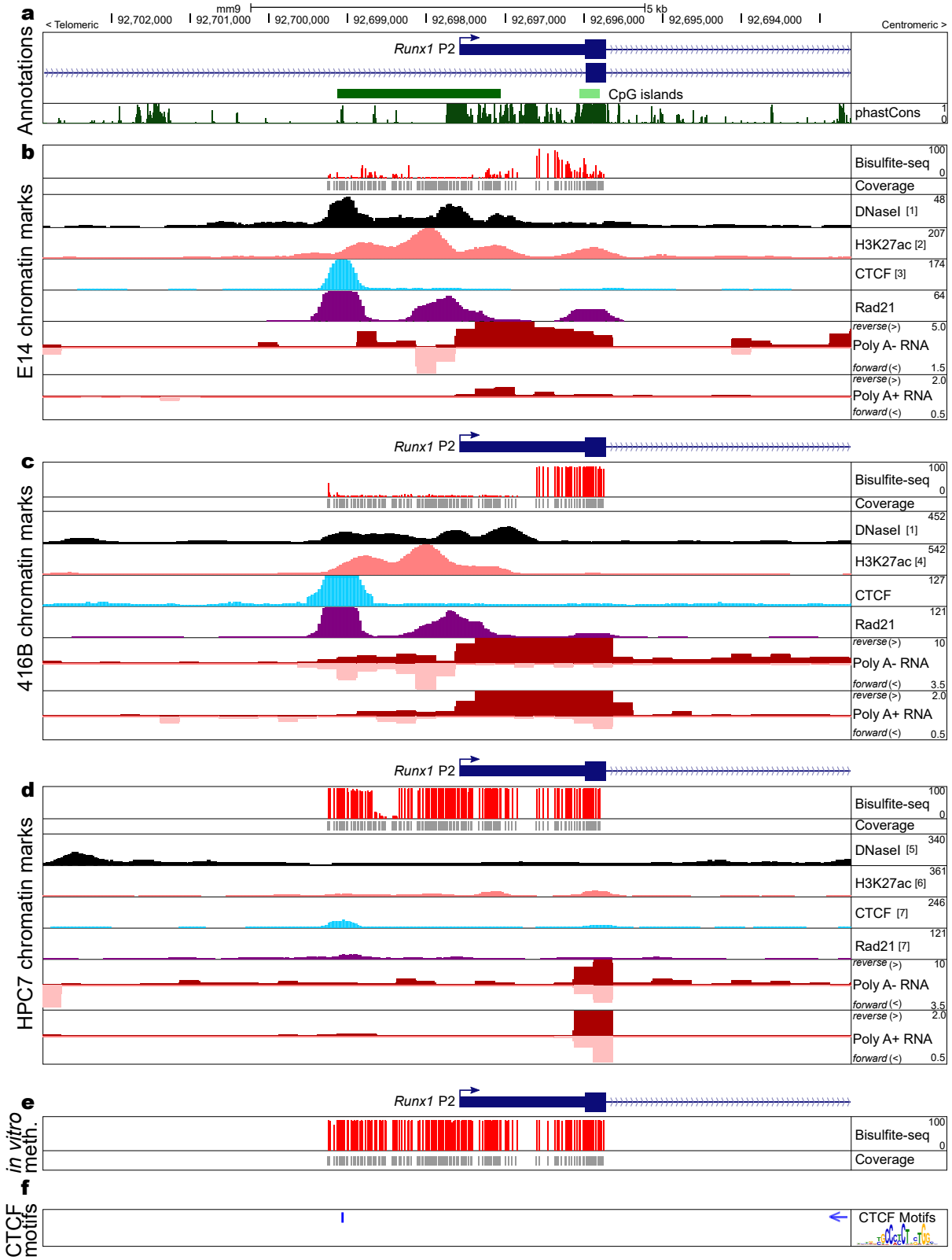


Figure 5.4 – CTCF binding and activity at the *Runx1* P2 is correlated with DNA methylation. *Legend continued on next page.*

Figure 5.4 – *Legend continued from previous page.* a) Annotation showing *Runx1* gene, P2 promoter, CpG island annotation from UCSC, and vertebrate conservation (phastCons). b) Chromatin marks in undifferentiated E14 mESCs including DNaseI ([1], Vierstra et al. 2014) H3K27ac ([2], Wamstad et al. 2012), CTCF ([3], Handoko et al. 2011), Rad21, poly A minus and plus RNA-seq. Targeted bisulfite sequencing is shown as red bars from 0 to 100 representing the percentage of methylated CpG dinucleotides at the population level. CpG dinucleotides covered are shown by grey bars (coverage). c) Chromatin marks in 416B myeloid progenitor cells including DNaseI ([1], Vierstra et al. 2014) H3K27ac ([4], Schütte et al. 2016), CTCF, Rad21, poly A minus and plus RNA-seq. Targeted bisulfite sequencing and coverage is shown. d) Chromatin marks in HPC7 haematopoietic progenitor cells including DNaseI ([5], Wilson et al. 2010a) H3K27ac ([6], Calero-Nieto et al. 2014), CTCF ([6], Calero-Nieto et al. 2014), Rad21 ([7], Wilson et al. 2016), poly A minus, and plus RNA-seq. The same scales as in Figure 5.4 were used to allow comparison between P1 and P2. Targeted bisulfite sequencing and coverage is shown. e) Targeted bisulfite sequencing of *in vitro* methylated control DNA shows near complete methylation. f) The orientation of *de novo* annotated CTCF motifs are shown in blue. UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_Methylation_P2`

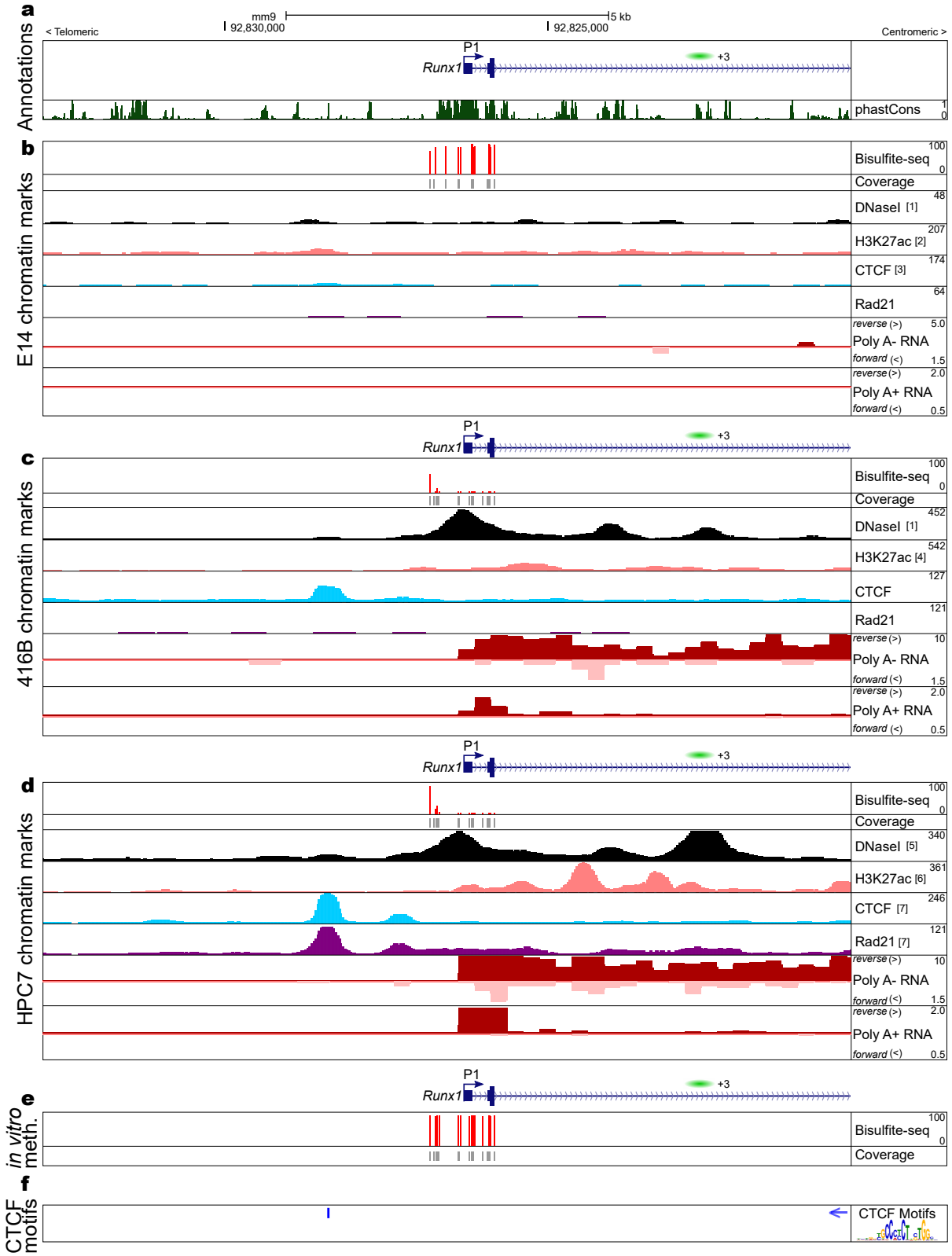


Figure 5.5 – CTCF binding and activity at the *Runx1* P1 is correlated with DNA methylation. *Legend continued on next page.*

Figure 5.5 – *Legend continued from previous page.* a) Annotation showing *Runx1* gene, P1 promoter, +3 enhancer that was previously shown to produce reporter expression in haematopoietic sites in developing mouse emrbyos (Schütte et al., 2016), CpG island annotation from UCSC, and vertebrate conservation (phastCons). b) Chromatin marks in undifferentiated E14 mESCs including DNaseI ([1], Vierstra et al. 2014) H3K27ac ([2], Wamstad et al. 2012), CTCF ([3], Handoko et al. 2011), Rad21, poly A minus and plus RNA-seq. The same scales as in Figure 5.4 were used to allow comparison between P1 and P2. Targeted bisulfite sequencing is shown as red bars from 0 to 100 representing the percentage of methylated CpG dinucleotides at the population level. CpG dinucleotides covered are shown by grey bars (coverage). c) Chromatin marks in 416B myeloid progenitor cells including DNaseI ([1], Vierstra et al. 2014) H3K27ac ([4], Schütte et al. 2016), CTCF, Rad21, poly A minus and plus RNA-seq. Targeted bisulfite sequencing and coverage is shown. d) Chromatin marks in HPC7 haematopoietic progenitor cells including DNaseI ([5], Wilson et al. 2010a) H3K27ac ([6], Calero-Nieto et al. 2014), CTCF ([6], Calero-Nieto et al. 2014), Rad21 ([7], Wilson et al. 2016), poly A minus, and plus RNA-seq. e) Targeted bisulfite sequencing of *in vitro* methylated control DNA shows near complete methylation. f) The orientation of *de novo* annotated CTCF motifs are shown in blue. Primer design was done by Danuta Jeziorska and bisulfite sequencing was done by Akin Bucakci and myself. UCSC session of these data: `http://genome-euro.ucsc.edu/s/dowens/allData_DO_Methylation_P1`

5.2.4 CTCF binding is correlated with promoter activity genome wide

The relationship between CTCF binding and promoter activity was examined genome-wide to explore whether the correlation seen at both *Runx1* promoters might also be applicable to other genes. Interestingly, CTCF peaks were strongly enriched at CGIs, promoters and 5' UTRs in all three cell types (Figure 5.6 a). Moreover, 16.7-25.4% of CTCF sites were within 5 kb of annotated TSSs between the three cell types (Figure 5.6 b). This showed that a significant number of CTCF peaks bound at or close to a TSS and could be involved in promoter regulation. Next, we asked the converse question of how many TSSs were bound by CTCF in each cell type. To make it possible to assign each CTCF peak to a TSS unambiguously, non-overlapping (single) TSSs with no other TSS within 5 kb were considered. This left 17263 out of 21396 (80.6%) of total TSSs in the refGene database. Out of non-overlapping TSSs, 5167 to 9986 bound CTCF somewhere within a 5 kb window centred on each TSS. On the other hand, 7277 to 12096 TSSs were not bound within a 5 kb window (Figure 5.6 c). A significant number of TSSs were bound by CTCF directly (within a 250 bp window centred on annotated TSSs), while many were associated with a more distal CTCF (Figure 5.6 c). In general, genes without CTCF bound were shared between the cell types, while genes with CTCF binding overlapped less well (Appendix Figure 7.19). This could imply that cell type-specific genes tend to be bound by CTCF in a cell type-specific manner, while housekeeping genes tend not to bind CTCF. In agreement with this, one third of CTCF peaks were unique to each cell type (Figure 5.6 a), showing that global CTCF binding patterns differ substantially between the cell types and could facilitate tissue-specific promoter regulation.

Interestingly, TSSs defined as active in each cell type (RPKM > 10th percentile of all TSS in each cell type) were highly enriched for CTCF binding compared to inactive TSS (Figure 5.6 e, χ^2 test, $p < 1\times10^{-10}$). Within a 5 kb window centred on single TSSs, CTCF binding strength (as measured by CTCF ChIP-seq read count or CTCF peak number) were both weakly positively correlated with TSS activity levels as measured by RNA-seq (Figure 5.6 f). Promoters with CTCF bound closer to the TSS (within 250 bp/2 kb/5 kb window centred on each TSS) showed increased H3K27ac enrichment and transcription levels (Figure 5.6 g, three-way ANOVA interaction effect [$F(1,12) = 16.3$, $p = 2.9\times10^{-35}$, $\eta^2 = 0.00049$] with Tukey's post-hoc test, *, adjusted $p < 3\times10^{-9}$). Together this suggests there is a genome-wide association between CTCF binding at promoters and marks of promoter activity. An explanation for this could be that open chromatin at active promoters leads to non-specific binding of CTCF. However, interestingly, a tendency (although not exclusivity) for CTCF to bind upstream of the TSS was observed for genes on both strands in all three cell types (Figures 5.7, 5.8, 5.9). Moreover, CTCF binding motifs showed a similar pattern of enrichment upstream of the TSS, suggesting that CTCF binds specifically to its motif rather than non-specifically at sites of open chromatin. CTCF motifs were absent from promoter regions that did not bind CTCF (Figures 5.7, 5.8, 5.9), suggesting that CTCF will likely bind its motif if present close to a TSS. TSS with CTCF bound were enriched for CGIs compared to TSS that did not bind CTCF.

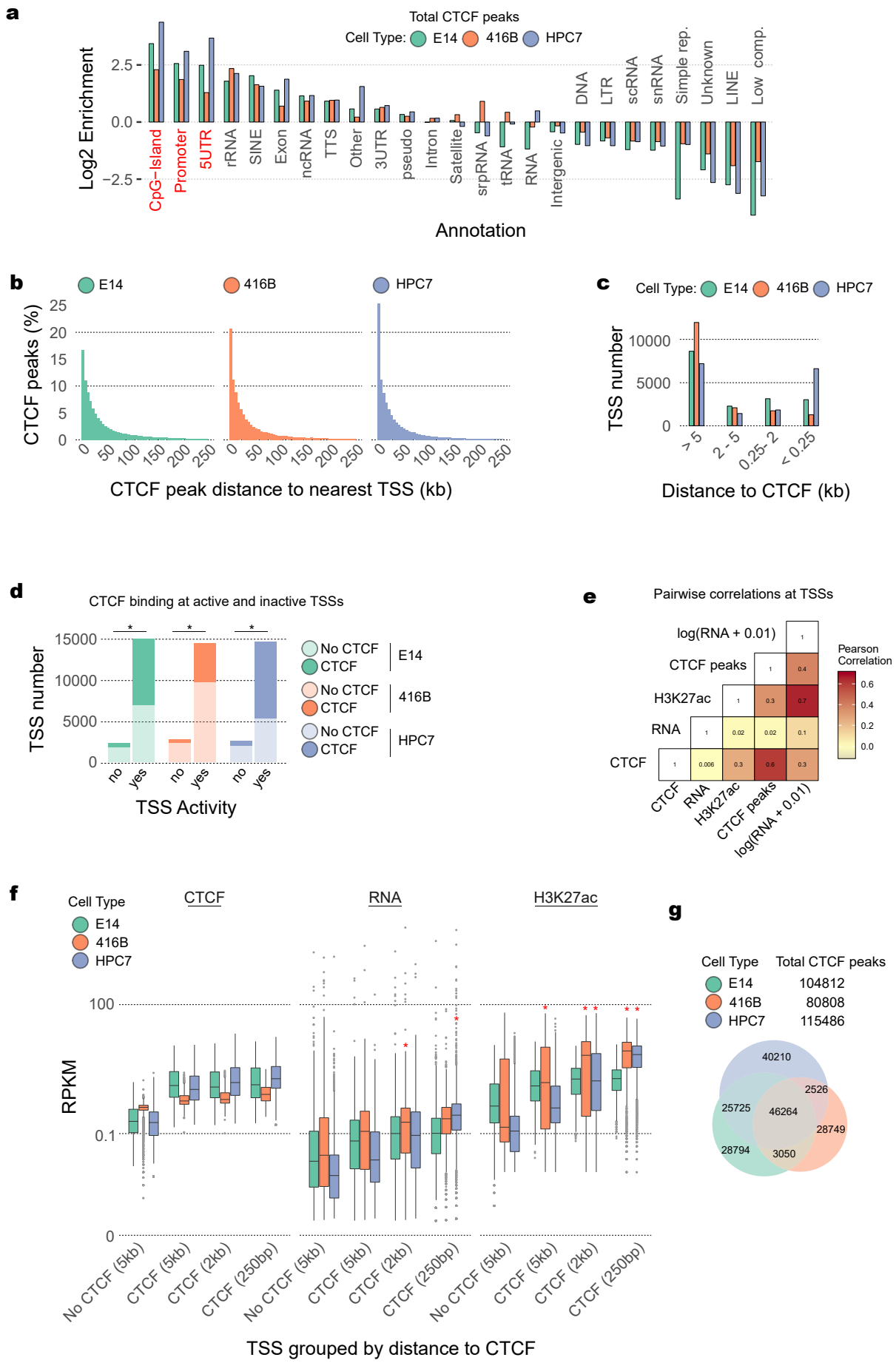


Figure 5.6 – CTCF binding is correlated with promoter activity genome wide. *Legend continued on next page.*

Figure 5.6 – *Legend continued from previous page.* a) Enrichment of total CTCF peaks at different genomic regions annotated using HOMER (Heinz et al., 2010). b) Distribution of total CTCF peaks by distance to the nearest annotated TSS in 5 kb bins. c) Number of non-overlapping TSS in each cell type grouped by distance to CTCF peaks. TSS were grouped based on whether CTCF bound within a 250 bp/2 kb/5 kb bin centered on each TSS, or did not bind within a 5 kb bin. d) Number of non-overlapping TSS with RPKM reads measured in a 5 kb bin centered on each TSS greater than the 10th percentile of expression in each cell type (active) or less than the 10th percentile of expression in each cell type (not active). The number of TSS in the active and not active groups that bound CTCF somewhere within a 5 kb window centred on each TSS is shown. χ^2 test, *, $p < 1{\times}10^{-10}$. e) Pairwise Pearson correlation matrix between normalised read counts (RPKM) quantified over a 5 kb window centred on single TSS for the different chromatin marks. RNA read counts were log normalised (+0.01) to better account for the large variation in read counts. f) Boxplot of CTCF ChIP-seq, poly A minus RNA-seq, and H3K27ac ChIP-seq in E14, 416B and HPC7 cells over non-overlapping TSSs grouped by distance to CTCF peaks. Three-way ANOVA with Tukey's post-hoc test, *, adjusted $p < 3{\times}10^{-9}$. Public ChIP-seq analysed were E14 CTCF ChIP-seq (Handoko et al., 2011) and H3K27ac ChIP-seq (Wamstad et al., 2012), HPC7 CTCF and H3K27ac ChIP-seq (Calero-Nieto et al., 2014). g) Overlap of all CTCF peaks called in undifferentiated E14 mESCs, 416B myeloid progenitor cells, and HPC7 haematopoietic progenitor cells.

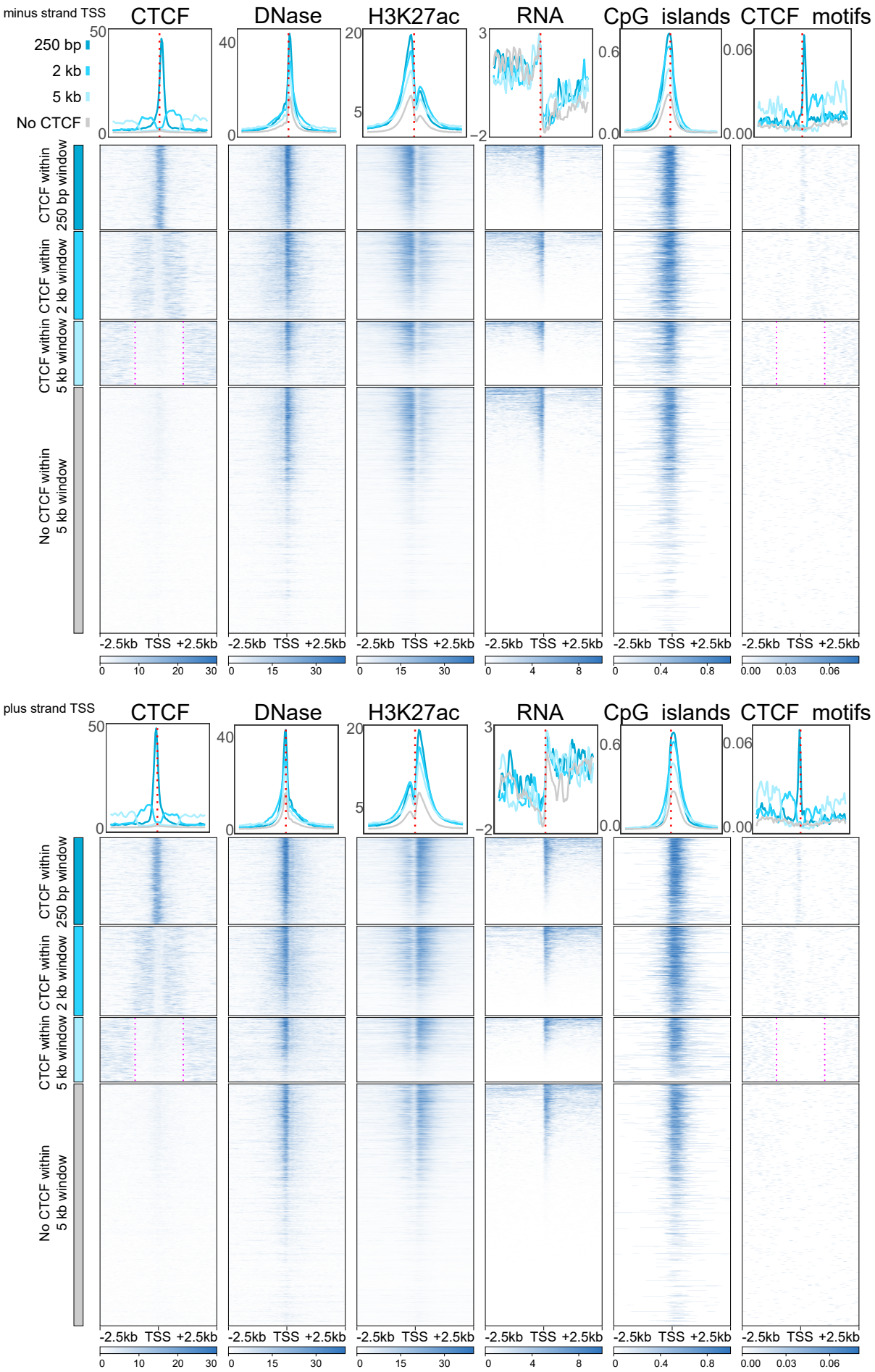


Figure 5.7 – Genome-wide meta-plot analysis of chromatin marks at TSS in E14 mESCs. *Legend continued on next page.*

Figure 5.7 – *Legend continued from previous page.* Meta-plot heatmaps and line plots of chromatin marks in E14 cells. Enrichment was calculated over a 5 kb window centred on TSS grouped by distance to CTCF as indicated by grey and shades of blue bars next to the heatmaps and the colour of the lines in the line plots. Minus strand genes are shown in the top set of plots and plus strand TSS are shown in the bottom set of plots. Pink dashed lines demarcate the boundary of regions depleted of CTCF binding and CTCF motifs. TSS were sorted on RNA-seq expression. CTCF, DNase, H3K27ac are shown as normalised RPKM values. Poly A plus RNA seq is shown as log(RPKM + 0.01). CpG islands and CTCF motifs are shown as relative arbitrary scores within a maximum possible range of zero to one. Public data analysed were E14 CTCF ChIP-seq (Handoko et al., 2011), DNaseI-seq (Vierstra et al., 2014), and H3K27ac ChIP-seq (Wamstad et al., 2012). CpG island annotations were taken from the UCSC browser. Enrichment of *de novo* annotated CTCF motifs derived from 416B CTCF ChIP-seq are shown.

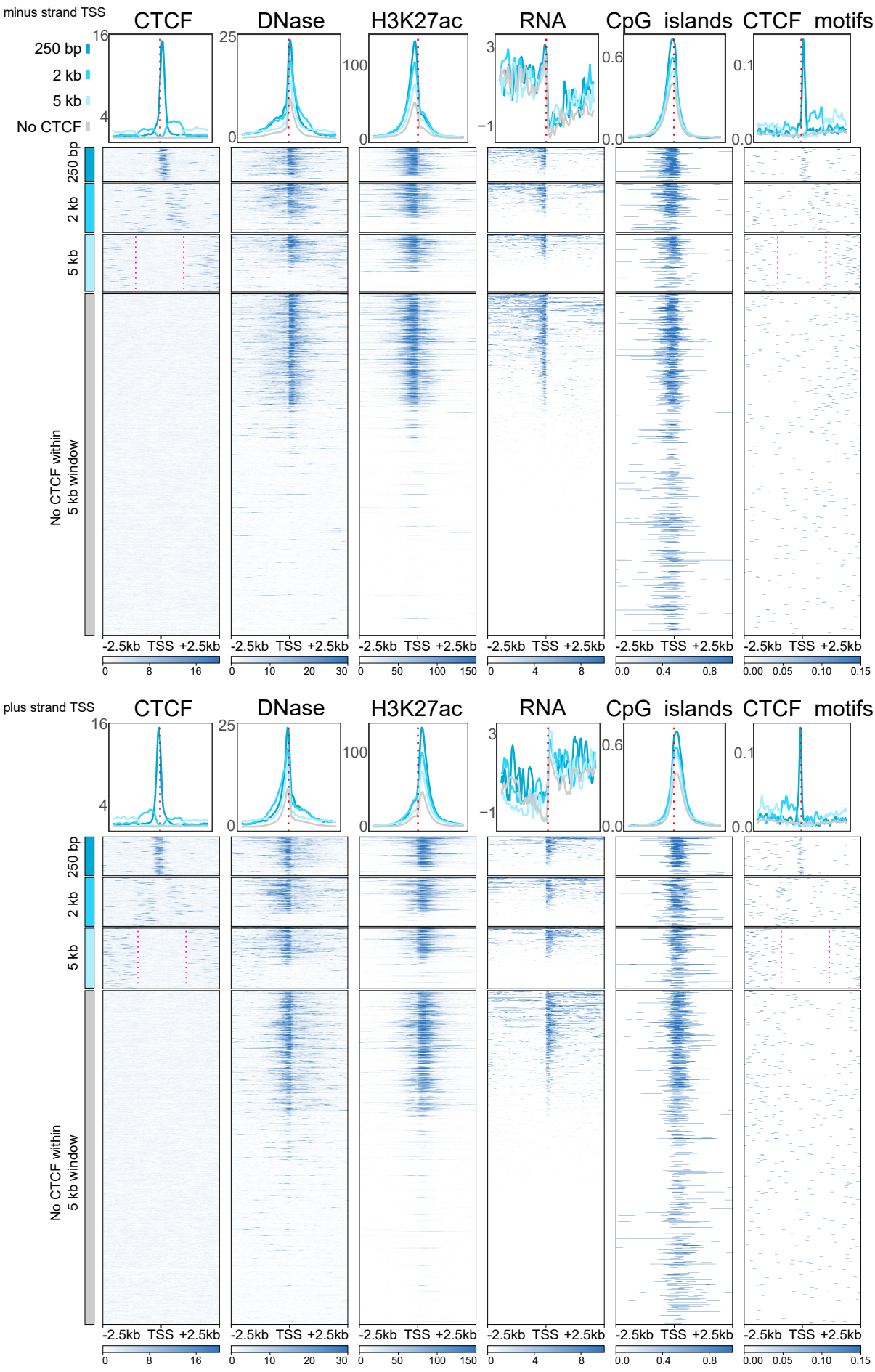


Figure 5.8 – Genome-wide meta-plot analysis of chromatin marks at TSS in 416B cells. *Legend continued on next page.*

Figure 5.8 – *Legend continued from previous page.* Meta-plot heatmaps and line plots of chromatin marks in 416B cells. Enrichment was calculated over a 5 kb window centred on TSS grouped by distance to CTCF as indicated by grey and shades of blue bars next to the heatmaps and the colour of the lines in the line plots. Minus strand genes are shown in the top set of plots and plus strand TSS are shown in the bottom set of plots. Pink dashed lines demarcate the boundary of regions depleted of CTCF binding and CTCF motifs. TSS were sorted on RNA-seq expression. CTCF, DNase, H3K27ac are shown as normalised RPKM values. Poly A plus RNA seq is shown as log(RPKM + 0.01). CpG islands and CTCF motifs are shown as relative arbitrary scores within a maximum possible range of zero to one. Public data analysed were 416B DNaseI-seq (Vierstra et al., 2014) and H3K27ac ChIP-seq (Schütte et al., 2016). CpG island annotations were taken from the UCSC browser. Enrichment of *de novo* annotated CTCF motifs derived from 416B CTCF ChIP-seq are shown.

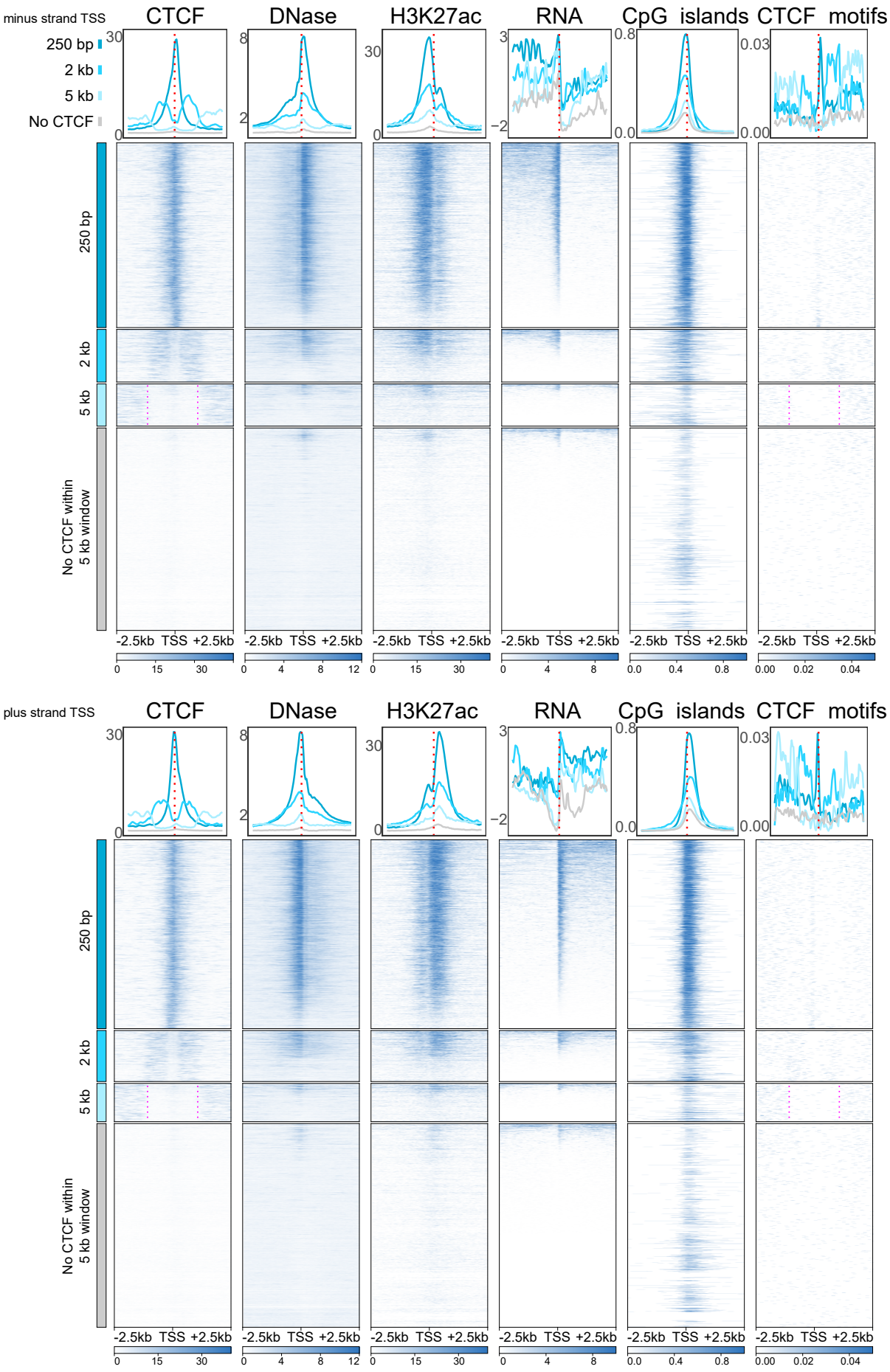


Figure 5.9 – Genome-wide meta-plot analysis of chromatin marks at TSS in HPC7 cells. *Legend continued on next page.*

Figure 5.9 – *Legend continued from previous page.* Meta-plot heatmaps and line plots of chromatin marks in HPC7 cells. Enrichment was calculated over a 5 kb window centred on TSS grouped by distance to CTCF as indicated by grey and shades of blue bars next to the heatmaps and the colour of the lines in the line plots. Minus strand genes are shown in the top set of plots and plus strand TSS are shown in the bottom set of plots. Pink dashed lines demarcate the boundary of regions depleted of CTCF binding and CTCF motifs. TSS were sorted on RNA-seq expression. CTCF, DNase, H3K27ac are shown as normalised RPKM values. Poly A plus RNA seq is shown as log(RPKM + 0.01). CpG islands and CTCF motifs are shown as relative arbitrary scores within a maximum possible range of zero to one. Public data analysed were HPC7 CTCF ChIP-seq (Calero-Nieto et al., 2014), DNaseI-seq (Wilson et al., 2010a), H3K27ac ChIP-seq (Calero-Nieto et al., 2014). CpG island annotations were taken from the UCSC browser. Enrichment of *de novo* annotated CTCF motifs derived from 416B CTCF ChIP-seq are shown.

5.2.5 Deletion of CTCF sites upstream of the *Runx1* P1 and P2 promoters using CRISPR/Cas9 in mouse embryonic stem cells

To facilitate the analysis of a possible causal relationship between CTCF binding at activity at *Runx1* promoters, CTCF sites upstream of *Runx1* promoters were targeted with Cas9 nuclease in E14 mESCs to generate mutants (Figures 5.10 a, 5.11 a). The rationale behind this would be to test differentiate mESC lines lacking CTCF sites upstream of the *Runx1* promoters to haematopoietic cells *in vitro* and then examine the role of these CTCF sites in alternative *Runx1* promoter usage. CTCF motifs were identified using *de novo* motif calling, and showed DNaseI footprints in 416B cells, and deep evolutionary conservation (Figures 5.10 b, 5.11 b). Short-range (S-R) PCR and Sanger sequencing identified two (ΔP1-CTCF) and three (ΔP2-CTCF) clones with unique deletions that abolished the predicted CTCF binding sites on both alleles (Figures 5.10 c and d, 5.11 c and d). These clones were then sub-cloned to reduce the possibility of the clones being mixed and harbouring residual wild type cells to yield a total of six ΔP1-CTCF clones and seven ΔP2-CTCF clones. Based on the possibility of Cas9-induced larger deletions (LDs) residing within the clones, a genotyping strategy involving successive PCRs up to 5 kb in length were performed (Figures 5.10 a, 5.11 a). In each of the clones, only expected band sizes were detected, implying that LDs within the range of the PCRs were not present. Furthermore, copy counting revealed that they all harboured the expected 2n copy numbers in the region up and downstream of the sgRNA target sites (Figures 5.10 e, 5.11 e). Therefore, these clones are unlikely to harbour LDs and will be useful tools to dissect the role of CTCF sites upstream of both *Runx1* promoters.

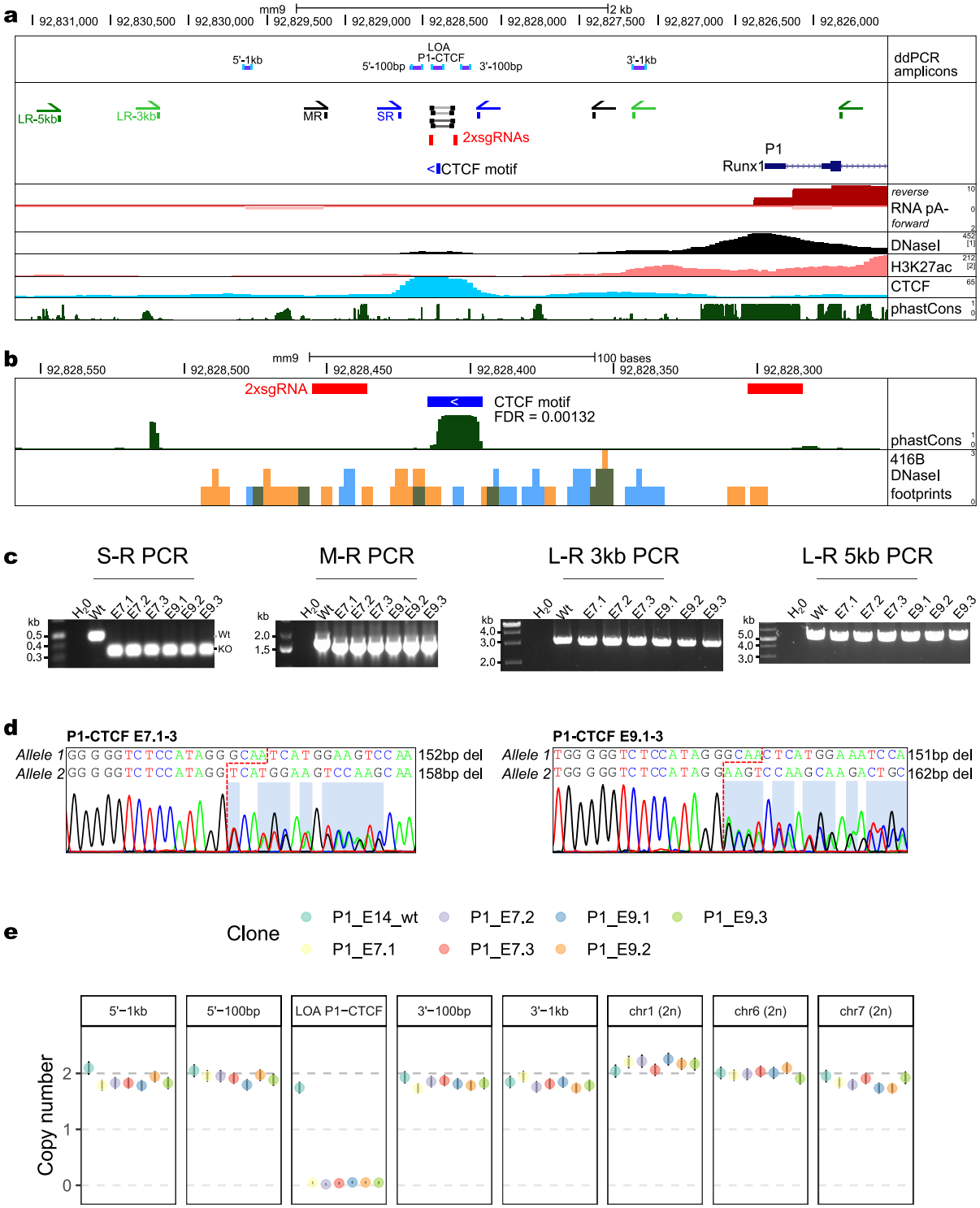


Figure 5.10 – Cas9 nucleasecan be used to examine requirements of CTCF binding sites upstream of the *Runx1* P1 promoter. *Legend continued on next page.*

Figure 5.10 – *Legend continued from previous page.* a) Locus map of the CTCF binding site upstream of the *Runx1* P1 promoter that was targeted in E14-TG2a mESCs using Cas9 nuclease and the indicated sgRNAs (red boxes). Four resolved deletions identified by Sanger sequencing of two separate clones are shown. Genotyping PCR primers and digital droplet (ddPCR) amplicons are shown. Public data analysed were 416B DNaseI-seq ([1], Vierstra et al. 2014) and H3K27ac ChIP-seq ([2], Vierstra et al. 2014). Poly A minus RNA-seq and CTCF ChIP-seq in 416B cells are also shown with phastCons 100 vertebrate conservation. b) Close up identification of the precise CTCF binding site showing conservation (phastCons 100 vertebrate), *de novo* annotated CTCF motif with FDR, DNaseI footprints in 416B cells, and sgRNAs used. c) Gel images show PCR amplification from gDNA of wild type (wt) E14-TG2a mESCs and knock-out clones with short-range (S-R) primers, medium-range (M-R) primers, 3 kb long-range (L-R 3 kb) and 5 kb long-range (L-R 5 kb) primers. Wt next to the gel image indicates the size of the wild type allele and KO indicates the size of alleles harbouring the expected deletion. d) Sanger sequencing of two separate clones (E7 and E9) that harboured heterozygous deletions abolishing the predicted CTCF binding site. Deletion lengths are shown in bp. Exact locations of the deletion breakpoint on each allele is shown as a dotted red line. e) ddPCR quantification of deletions across a 2.2 kb window centred on P1-CTCF binding site. Each point represents the mean relative copy number of the target DNA sequence (+/- 95% confidence interval) of a single clone according to the colour key.

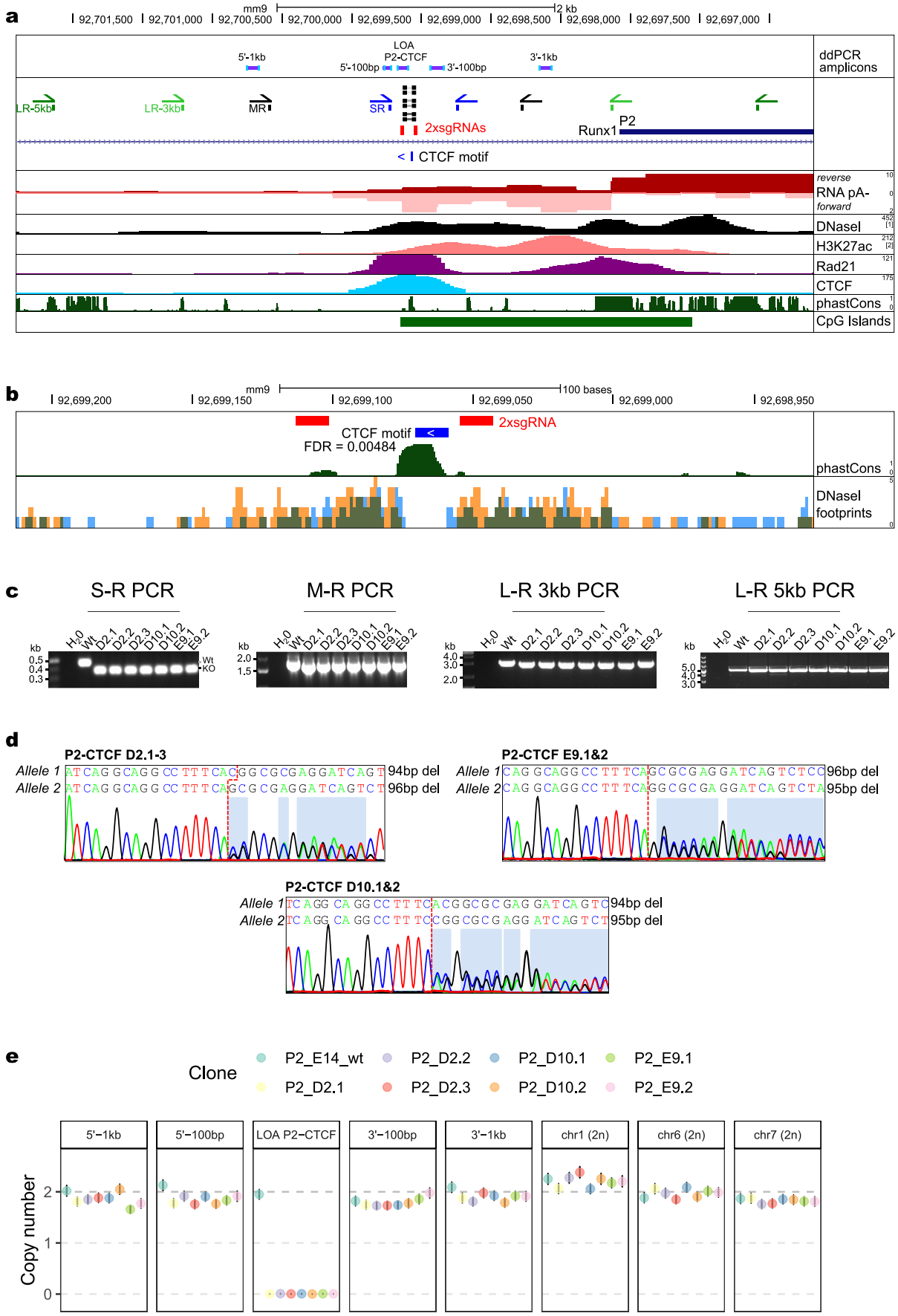


Figure 5.11 – Cas9 nuclease can be used to examine requirements of CTCF binding sites upstream of the *Runx1* P2 promoter. *Legend continued on next page.*

Figure 5.11 – *Legend continued from previous page.* a) Locus map of the CTCF binding site upstream of the *Runx1* P2 promoter that was targeted in E14-TG2a mESCs using Cas9 nuclease and the indicated sgRNAs (red boxes). Six resolved deletions identified by Sanger sequencing of three separate clones are shown. Genotyping PCR primers and digital droplet (ddPCR) amplicons are shown. Public data analysed were 416B DNaseI-seq ([1], Vierstra et al. 2014) and H3K27ac ChIP-seq ([2], Vierstra et al. 2014). Poly A minus RNA-seq, CTCF ChIP-seq, and Rad21 ChIP-seq in 416B cells are also shown with phastCons 100 vertebrate conservation. b) Close up identification of the precise CTCF binding site showing conservation (phastCons 100 vertebrate), *de novo* annotated CTCF motif with FDR, DNaseI footprints in 416B cells, and sgRNAs used. c) Gel images show PCR amplification from gDNA of wild type (wt) E14-TG2a mESCs and knock-out clones with short-range (S-R) primers, medium-range (M-R) primers, 3 kb long-range (L-R 3 kb) and 5 kb long-range (L-R 5 kb) primers. Wt next to the gel image indicates the size of the wild type allele and KO indicates the size of alleles harbouring the expected deletion. d) Sanger sequencing of three separate clones (D2, D10, and E9) that harboured heterozygous deletions abolishing the predicted CTCF binding site. Deletion lengths are shown in bp. Exact locations of the deletion breakpoint on each allele is shown as a dotted red line. e) ddPCR quantification of deletions across a 2.2 kb window centred on P2-CTCF binding site. Each point represents the mean relative copy number of the target DNA sequence (+/- 95% confidence interval) of a single clone according to the colour key.

5.3 Chapter conclusions and discussion

In this chapter, I examined a possible role for CTCF/cohesin in *Runx1* alternative promoter usage. Previous studies implicated both factors in this process (Horsfield et al., 2007; Marsman et al., 2014; Nora et al., 2017), but whether *Runx1* is directly or indirectly regulated had not previously been examined. CTCF and Rad21 binding in the *Runx1* locus was examined in E14, 416B, and HPC7 cells. CTCF bound at multiple sites in common between the cell types, particularly at the boundaries of the *Runx1* regulatory domain. Notably, CTCF sites at the *Runx1* promoters differed from the boundary CTCF sites in that they bound in a tissue-specific manner. Previously, tissue-invariant CTCF sites were shown to be enriched at TAD boundaries defined by Hi-C, while genome-wide, tissue-specific CTCF sites were predominantly found within TADs and not at their boundaries (Van Bortle et al., 2014). Together this provides support for the notion that tissue-specific CTCF sites at the *Runx1* promoters might be performing regulatory functions alongside (or instead of) the well-described boundary formation properties of CTCF (Phillips and Corces, 2009). CTCF binding was previously seen at the *runx1/RUNX1* promoters (Marsman et al., 2014; Schuijers et al., 2018), but to the best of my knowledge cell type-specific binding at these sites has not been reported before. At present, it cannot be excluded that some of the observed differences in CTCF binding by ChIP-seq is due to the fact that the experiments were performed at different times or using different conditions. They were all done using the same polyclonal antibody, but using different batches. Additional experiments using three replicates of ChIP performed in parallel would provide the best statistical comparison between enrichment in the different cell types.

In vitro assays suggested that both tissue-specific and tissue-invariant CTCF sites act as insulators in the context of episomal plasmids. However, several caveats make it difficult to conclude that these CTCF sites exhibit the same insulating properties in their endogenous chromatin environment. Firstly, episomal plasmids do not integrate into the genome and will be chromatinised differently to the endogenous locus (Riu et al., 2007). Secondly, the plasmids used were circular and therefore might exhibit different looping properties to linear DNA. Linearised insulator plasmids were tested but yielded low transfection efficiencies and unreliable luciferase expression levels (data not shown). The fact that mutating or deleting predicted CTCF binding motifs did not reduce the insulating properties of the CTCF sites suggests that their insulating properties were not CTCF-dependent. An alternative explanation for this could be that any sequence cloned into the constructs would have had a non-specific repressive effect. To exclude this possibility, the insulating properties of several different genomic DNA sequences without predicted CTCF binding motifs or other regulatory activity should be examined.

Another possible cause of the CTCF-independent repressive effects could be mediated by DNA methylation of the exogenous plasmid sequences. It is likely that they would be methylated upon entry into the nucleus because the ∼500 bp cloned CTCF sites were CpG rich (-468-CTCF had 55% GC content and P2-CTCF had 63% GC content). This methylation could then spread to the nearby enhancer and promoter, repressing luciferase expression, as DNA methylation is associated with transcriptional silencing of promoters (Curradi et al., 2002). It was interesting that mutating the P2-CTCF motif reduced luciferase activity compared to wild type P2-CTCF because this would

be consistent with CTCF binding to the P2 CGI normally inhibiting DNA methylation. Indeed, it has previously been suggested that CTCF binding could inhibit DNA methylation (Szabo et al., 2004; Davalos-Salas et al., 2011). Interestingly, many repeat elements carry CTCF motifs (Bourque et al., 2008), which were proposed to have been selected for as a mechanism to inhibit DNA methylation-mediated silencing (Merkenschlager and Odom, 2013). In further support of CTCF binding inhibiting DNA methylation is the fact that CTCF paralog BORIS (also known as CTCFL), which shares its 11 zinc-finger domains with CTCF, is associated with the erasure of global DNA methylation marks during male gametogenesis (Loukinov et al., 2002; Vatolin et al., 2005). Therefore, it is possible that DNA methylation of the exogenous CTCF sites led to DNA methylation that repressed luciferase expression.

Runx1 promoters bound CTCF upstream of their TSS, and this was correlated with promoter activity. It remains to be established, however, whether CTCF binding is merely correlated with promoter activity, or a cause of it. CTCF binding at promoters might be causally related to promoter activation in several ways. Firstly, it has been shown that CTCF binding to promoters can facilitate enhancer-promoter interactions (Schuijers et al., 2018; Canzio et al., 2019; Hyle et al., 2019). In this way, CTCF binding close to the *Runx1* promoters might allow enhancers to engage with the promoter, increasing its transcriptional output. Second, since CTCF has been suggested to play a role in inhibiting DNA methylation (Szabo et al., 2004; Davalos-Salas et al., 2011), and we saw potential evidence of this in our *in vitro* luciferase assays, CTCF binding at the promoters could be inhibiting DNA methylation and thus preventing promoter silencing. Alternatively, CTCF binding at promoters has been suggested to prevent nucleosome occlusion (Nora et al., 2017), which might help maintain an open chromatin state permissive for transcription to occur. Therefore, several possible mechanisms exist for how CTCF binding could modulate promoter activity.

Several clues in our data suggest that the CTCF binding at *Runx1* promoters, and promoters genome-wide, might be causally related to their activity. Firstly, binding of CTCF occurred upstream of the TSS at *Runx1* promoters when they were active. Interestingly, there was a preference for this tendency genome-wide at TSSs on **either strand**. Second, promoters that bound CTCF were enriched for CTCF motifs. Since the CTCF binding motif is ∼20 bp in length and CTCF motifs upstream of both *Runx1* promoters (Figures 5.10 and 5.11) and also at the *MYC* promoter (Schuijers et al., 2018) are deeply evolutionarily conserved, this argues that they were selected for and do play an important regulatory function. Finally, up to one quarter of CTCF peaks were found within 5 kb of TSSs, and approximately one third of CTCF sites were unique to each cell type, suggesting that differential binding of CTCF sites close to promoters could be playing a role in establishing cell type-specific promoter regulation.

Alternative explanations for the observed enrichment of CTCF binding close to active promoters should also be considered. One is that transcriptionally silent promoters are often marked by DNA methylation (Curradi et al., 2002). Since CTCF binding has been suggested to be inhibited by DNA methylation (Bell and Felsenfeld, 2000; Shukla et al., 2011; Wang et al., 2012; Flavahan et al., 2016; Hashimoto et al., 2017;

Xu and Corces, 2018; Schuijers et al., 2018; Canzio et al., 2019), it might be expected that CTCF would not bind these methylated and silent promoters. While methylated and silenced promoters would be expected to not bind CTCF, this does not explain why CTCF did bind at some active promoters, unless CTCF simply binds at its motif whenever it is present in accessible and unmethylated DNA, without necessarily playing a role in promoter activity. However, since CTCF motifs were deeply evolutionarily conserved, this seems unlikely. An explanation for why CTCF might bind preferentially upstream of a TSS could be that CTCF binding downstream of a TSS might be inhibitory to transcription so may be selected against during evolution. Indeed, CTCF has been suggested to slow down pol II progression (Shukla et al., 2011). However, even if CTCF binding was only observed downstream of the TSS, downstream elements such as the downstream promoter element (DPE) are still capable of regulating transcription (Kadonaga, 2002). Therefore, the observed enrichment of CTCF binding and CTCF motifs makes it seem plausible that CTCF plays a direct role in regulating a subset of cell type-specific genes, including *Runx1*. In order to directly test the causal role of CTCF binding to the *Runx1* promoters, CTCF binding motifs upstream of each promoter were deleted in mESCs. Unfortunately, initial CTCF knock-out clones harboured Cas9-induced large deletions (Owens et al., 2019) and new clones had to be generated. These have now been fully genotyped and are awaiting further experiments which are discussed in section 6.1.6.

We showed that differential CTCF binding upstream of the *Runx1* P1 and P2 promoters was correlated with promoter DNA methylation status. It was interesting that genome-wide, promoters that bound CTCF were enriched for CGIs in all three cell types examined. This suggests that DNA methylation could be a pervasive mechanism to modulate CTCF binding to promoters, since CGIs are classically associated with regulation by DNA methylation (Deaton and Bird, 2011). DNA methylation was examined over the *Runx1* P2 CTCF binding site but the ∼1.8 kb distance meant that methylation of the CTCF binding site upstream of P1 was not directly measured. Bisulfite sequencing of the P1 CTCF binding site would show whether methylation at the promoter region extends to the CTCF motif. In addition, whole-genome bisulfite sequencing could be added to our genome-wide analyses to correlate CTCF binding, promoter activity, and DNA methylation status. Genome-wide, most methylation sensitive CTCF sites and most tissue-specific CTCF sites are located within TADs (Van Bortle et al., 2014; Wang et al., 2012; Maurano et al., 2015), agreeing that methylation could be a mechanism to establish cell-type specific CTCF binding patterns at tissue-specific gene promoters.

However, it remains unclear what could be regulating methylation at the *Runx1* promoters. It was previously shown that transcriptionally active promoters exclude DNA methyltransferases 3A and 3B and remain demethylated (Baubec et al., 2015). The P2 promoter was active in 416B cells and low-level transcription was seen in E14 mESCs, which might prevent DNA methylation of P2 in both of these cell types. Alternatively, CTCF binding upstream of the P2 promoter might prevent its methylation, as suggested by the insulator assays discussed above and in prior work (Szabo et al., 2004; Davalos-Salas et al., 2011). In the case of the *Runx1* P1 promoter, it was methylated in E14 mESCs and then became demethylated in 416B and HPC7 cells. Therefore, a mechanism must exist for demethylating this promoter. It is pos-

sible that P1 is demethylated when a specific enhancer becomes active that interacts with it. Biochemical incompatibility between different enhancers and promoters has been suggested to provide such enhancer-promoter specificity (Li and Noll, 1994; van Arensbergen et al., 2014). This is unlikely to be the case for the +23 enhancer (which is active from the HE stage onwards) as this was shown to upregulate both P1 and P2 promoter activity in transgenic assays (Bee et al., 2009a). CTCF-mediated chromatin boundaries may be more important for enhancer-promoter specificity (Lettice et al., 2011; Symmons et al., 2014; de Wit et al., 2015; Guo et al., 2015; Lupianez et al., 2015; Franke et al., 2016; Hanssen et al., 2017). Indeed, the P2 promoter exhibited a boundary effect, separating the gene into two compartments (see chapter 3, Figure 3.13), and so any of the enhancers upstream of P2 (+3, +23, +110) might predominantly activate P1.

Another cause of P1 demethylation could be non-coding anti-sense transcription occurring from the enhancers in the first intron of the gene (which was observed in haematopoietic cells by poly-A minus RNA-seq). In agreement with this, it was recently shown that anti-sense transcription was associated with DNA demethylation, CTCF recruitment and alternative promoter usage at the α-protocadherin locus (Canzio et al., 2019). Indeed, several studies have reported non-coding transcription-induced DNA demethylation at both CGI (Canzio et al., 2019; Arab et al., 2019) and non CGI regions (Benner et al., 2015; Isoda et al., 2017), possibly mediated by recruitment of TET enzymes (Canzio et al., 2019; Arab et al., 2019). In further support of a role for non-coding eRNAs in demethylating the P1 promoter, open chromatin peaks and TF binding were seen at +3 enhancer in HE cells before P1 promoter activity would be expected to increase (see chapter 4 Figure 4.4). Together this suggests that eRNA-induced DNA demethylation could be leading to demethylation and activation of the *Runx1* P1 promoter.

An alternative cause of demethylation at P1 could be a passive loss of methylation occurring over developmental time after successive rounds of DNA replication. This type of mechanism has been reported before, particularly at late DNA replicating regions (Shipony et al., 2014; Zhou et al., 2018). Analysis of the *Runx1* gene showed that the P1 promoter is indeed late replicating in mESCs and switches to early replicating in haematopoietic cells (data not shown) suggesting that this passive loss of methylation could be occurring at *Runx1*. Whatever the primary cause, demethylation of the P1 promoter might allow CTCF binding here leading to promoter activation.

Based on the correlations between CTCF binding, DNA methylation, and activity at the *Runx1* promoters, differential promoter methylation is a plausible regulatory mechanism that might regulate *Runx1* promoter choice during development. Indeed, methylation has previously been shown to transcriptionally control the promoters of other *Runx* family genes including the *Runx2* P1 promoter in mESCs and chrondrocytes (Fouse et al., 2008; Takahashi et al., 2017). As my study used mESCs and immortalised haematopoietic progenitor cell lines, it will be of interest to assess whether similar mechanisms are at play in cell types undergoing developmental haematopoiesis *in vivo*.

6. General Discussion

6.1 Summary of results

The overall aim of this project was to examine the *cis*-regulatory mechanisms of *Runx1* during developmental haematopoiesis. A better understanding of the regulatory mechanisms of *Runx1* extends current knowledge about how large and complex genes (with multiple promoters and enhancers) are regulated during development. Moreover, understanding mechanism of *Runx1* regulation could inform directed differentiation protocols for the *de novo* generation of haematopoietic stem cells (HSCs) *in vitro*, and provide potential mechanisms to manipulate *RUNX1* expression in human leukaemia.

Chapter 3 examined differences in *cis*-interactions between cells transcribing low and high levels of *Runx1*. The overall ∼1.1 Mb *Runx1* TAD was fully formed in undifferentiated mESCs that transcribed low levels of *Runx1*. Therefore, the formation of the *Runx1* TAD is likely independent of transcription. In both cell types, distinct sub-TADs were observed within the overall TAD which is consistent with a fractal organisation of the genome, consisting of nested interaction domains. Increased 'stripes' and interaction foci were observed in 416B cells expressing higher levels of *Runx1* compared to undifferentiated mESCs, indicative of increased processivity or rate of loop extrusion when transcription levels were high. Specific interactions were observed between the *Runx1* promoters and haematopoietic enhancers previously identified (Nottingham et al., 2007; Bee et al., 2009b; Schütte et al., 2016), arguing that they regulate *Runx1* transcription in haematopoietic cells. Long-range interactions were observed between *Runx1* and another highly expressed gene on the same chromosome (*Erg*), suggesting that a subset of *cis*-interactions are mediated by mechanisms other than loop extrusion, and possibly by phase separation and condensate formation.

In chapter 4 I focussed on redundancies between *Runx1* enhancers in the endogenous locus during EHT *in vitro*. The *Runx1* +23, +110, and +204 enhancers all showed dynamic changes in chromatin accessibility that agreed with the activities of the enhancers in transgenic enhancer-reporter studies previously done in the lab (Nottingham et al. 2007; Bee et al. 2009b; Swiers et al. 2013a; Schütte et al. 2016 and de Bruijn lab, unpublished observations). This suggests that these enhancers are important for upregulating *Runx1* expression during EHT. Conserved transcription factor (TF) binding sites (TFBS) that were unique to each enhancer were shown to be important for enhancer function *in vitro*. Combined with gene expression analysis, this suggested that some of these upstream TFs might also regulate *Runx1* enhancers during EHT. Deletion of *Runx1* enhancers was done to examine whether redundancy and/or synergy exists between them. Unfortunately, these studies were hampered by the occurrence of large deletions that thwarted our experiments. Detailed analysis of these deletions revealed that microhomologies were often seen at their breakpoints, suggesting that MMEJ likely plays a role in larger deletion (LD) generation. These findings were recently published (Owens et al., 2019). After estab-

lishing a robust genotyping strategy for CRISPR/Cas9-edited clones, correctly edited clones were identified and analysed. Deleting the +110 enhancer reduced Runx1 levels and impacted haematopoietic cell generation during *in vitro* EHT. On the other hand, deleting the +204 enhancer had no effect on *Runx1* levels or haematopoietic differentiation. This suggests that the enhancers play somewhat distinct roles during EHT. Therefore, it is likely that enhancer redundancy exists since deletion of a single enhancer only moderately impacted Runx1 levels. Moreover, since deleting a single enhancer did not drastically reduce *Runx1* levels, it seems likely that the haematopoietic *Runx1* enhancers do not act synergistically, arguing against the notion that they could be part of a super enhancer.

Chapter 5 examined a possible role for CTCF binding and DNA methylation in *Runx1* alternative promoter regulation. Constitutive and dynamic CTCF sites exhibited insulator properties in *in vitro* assays, suggesting that they might both perform similar barrier functions in endogenous chromatin. Cell type-specific CTCF binding upstream of the *Runx1* promoters was seen and this correlated with promoter activity. DNA methylation at the promoters was associated with reduced CTCF binding and promoter silencing. Together this implies that CTCF binding might be causally related to *Runx1* promoter activity. Genome-wide, active promoters were enriched for CTCF binding to motifs often upstream of their TSS, suggesting that modulating CTCF binding close to promoters might be a general mechanism for regulating promoter activity.

6.1.1 Differences in *cis*-interactions within the *Runx1* domain between transcriptionally inactive and active states

The *Runx1* TAD was established in mESCs before gene transcription was upregulated. This is consistent with a model where TADs are established early on during development prior to tissue-specific gene upregulation, possibly by a background level of loop extrusion (Hug et al., 2017; Brown et al., 2018; Oudelaar et al., 2019) (Figure 6.1). But if loop extrusion is happening everywhere all of the time, what changes to facilitate gene upregulation? *Cis*-interactions indicative of loop extrusion were strengthened in 416B cells (where *Runx1* transcription was high) compared to E14 mESCs where *Runx1* transcription was low. This suggests that loop extrusion may be involved in gene upregulation by establishing enhancer-promoter interactions. Cohesin complexes extruding chromatin loops will traverse the entire domain pervasively, meaning that all regions within the loop will be brought into proximity with each other after passing through a single cohesin complex, which might facilitate enhancer-promoter interactions (Figure 6.1). Since CTCF/cohesin is not required to maintain the majority of enhancer-promoter interactions (Nora et al., 2017; Rao et al., 2017; Wutz et al., 2017; Schwarzer et al., 2017; Hyle et al., 2019), other factors are likely involved in maintaining enhancer-promoter interactions. One possibility is that condensates of transcriptional activators such as Mediator, BRD4, pol II or tissue-specific TFs form at enhancers and promoters (Figure 6.2). After being brought into proximity by loop extrusion, condensates at enhancers and promoters could interact with each other, stabilising enhancer-promoter interactions and allowing productive transcription to occur. Importantly, multiple enhancers might be interacting together (Oudelaar et al., 2018) suggesting that large condensates could stabilise interactions

between *Runx1* promoters and multiple enhancers (Figure 6.2).

It is difficult to precisely pinpoint whether changes in loop extrusion led to the changes in transcription, or vice versa, as the increased levels of loop extrusion observed here were seen in two cell lines. It is also possible that changes in loop extrusion are necessary to counteract high levels of transcription, which has been suggested to generate torsional stress in DNA that can stall transcription (Ma et al., 2019). More refined time-course experiments are being planned where Tiled-C will be done at different stages of haematopoietic differentiation of mESCs *in vitro*. Undifferentiated mESCs will be used as the ground state with low *Runx1* transcription levels, Flk1+ mesoderm will be used as the intermediate state before the onset of high levels of *Runx1*, and CD41+ haematopoietic progenitor cells that express high levels of *Runx1* will be the final stage analysed. This time-course should allow the changes in *cis*-interactions in the *Runx1* domain to be examined more precisely over time, and should provide better insight into whether increased loop extrusion occurs prior to or after gene upregulation. Assuming that increased loop extrusion causes transcriptional upregulation, it is unclear what causes an increase in loop extrusion. One possibility is that enhancers may play a role in recruiting cohesin. If this was the case, then loss of an enhancer should decrease the recruitment of cohesin complexes, reduce the frequency of loop extrusion, and decrease transcription. This hypothesis could be tested by deleting *Runx1* enhancers in mESCs and performing Tiled-C on the same stages of differentiation outlined above.

6.1.2 Hierarchical organisation of the *Runx1* regulatory domain

A hierarchical organisation of sub-TAD structures were observed throughout the entire 1.1 Mb *Runx1* TAD and within the *Runx1* gene itself. Indeed, the genome is generally considered to be fractal in organisation (Bancaud and Ellenberg, 2012; Gibcus and Dekker, 2013; Oudelaar and Higgs, 2017), with multiple levels of nested interaction domains. Importantly, the subcompartmentalisation within the *Runx1* gene is different to smaller genes including α- and β-globin. In the case of the α-globin genes, they reside with their enhancers entirely within their own sub-TAD (Hay et al., 2016; Hanssen et al., 2017). Of note, this entire α-globin sub-TAD fits within the first intron of *Runx1*. It is possible that sub-TADs within *Runx1* are simply required due to its large size. However, because sub-TADs direct enhancer-promoter specificity (Hanssen et al., 2017), establishing multiple sub-TADs within the same gene might also allow for more complex transcriptional regulation. For example, the two promoters of *Runx1* could be regulated by different sets of enhancers. This notion is supported by the fact that P1 promoter interacted most strongly with +23 and +110 enhancers and less so with +204. However, it remains to be tested whether the sub-TADs within *Runx1* perform a functional role. One way this could be tested is by performing Tiled-C analysis in mESCs lacking CTCF sites upstream of the *Runx1* promoters. If the CTCF site played a role in sub-TAD formation then this would be seen as decreased boundary formation at the promoters. However, this analysis will be complicated by the fact that sub-TAD boundaries were correlated to transcriptional activity and CTCF binding upstream of the promoters. Therefore, if deleting CTCF sites upstream of the promoters altered their activity, it would not be clear whether this was mediated by impacting sub-TAD boundary formation or by

decreasing promoter activity through another mechanism. One way to shed further light on whether sub-TAD boundaries are tissue-specific would be to perform Tiled-C in HPC7 cells. The P2 promoter in these cells is methylated, inactive, and does not bind CTCF. Therefore, the sub-TAD boundary at P2 that was observed in both 416B and E14 cells might be expected to be absent in HPC7 cells, which would agree that tissue-specific sub-TADs form within the *Runx1* gene and are associated with changes in its transcriptional regulation.

6.1.3 Long-range interactions between highly transcriptionally active elements

Not all *cis*-interactions occurred within the *Runx1* regulatory domain. Long-range interactions were observed between *Runx1* enhancers and *Erg*, located 3 Mb away on mouse chromosome 16. Phase separation of transcriptionally active components is a plausible candidate for driving these interactions. Both regions were annotated as SEs and SEs have previously been suggested to form higher order condensates, possibly mediated by phase separation (Cho et al., 2018). It remains unclear whether long-range interactions between *Runx1* and *Erg* played a functional role in transcriptional regulation of either gene. Perturbing the interaction between *Runx1* enhancers and *Erg* (by gene editing cells such that *Erg* and *Runx1* reside on different chromosomes, for example) could help elucidate whether this interaction is dependent purely on phase separation of transcriptionally active components or whether linear genomic distance also plays a role. This could help elucidate the relative contribution of loop extrusion or phase separation in the long-range interaction between *Runx1* and *Erg*. Moreover, it would be possible to determine if these long-range interactions were functional if interactions were reduced by placing the genes on different chromosomes.

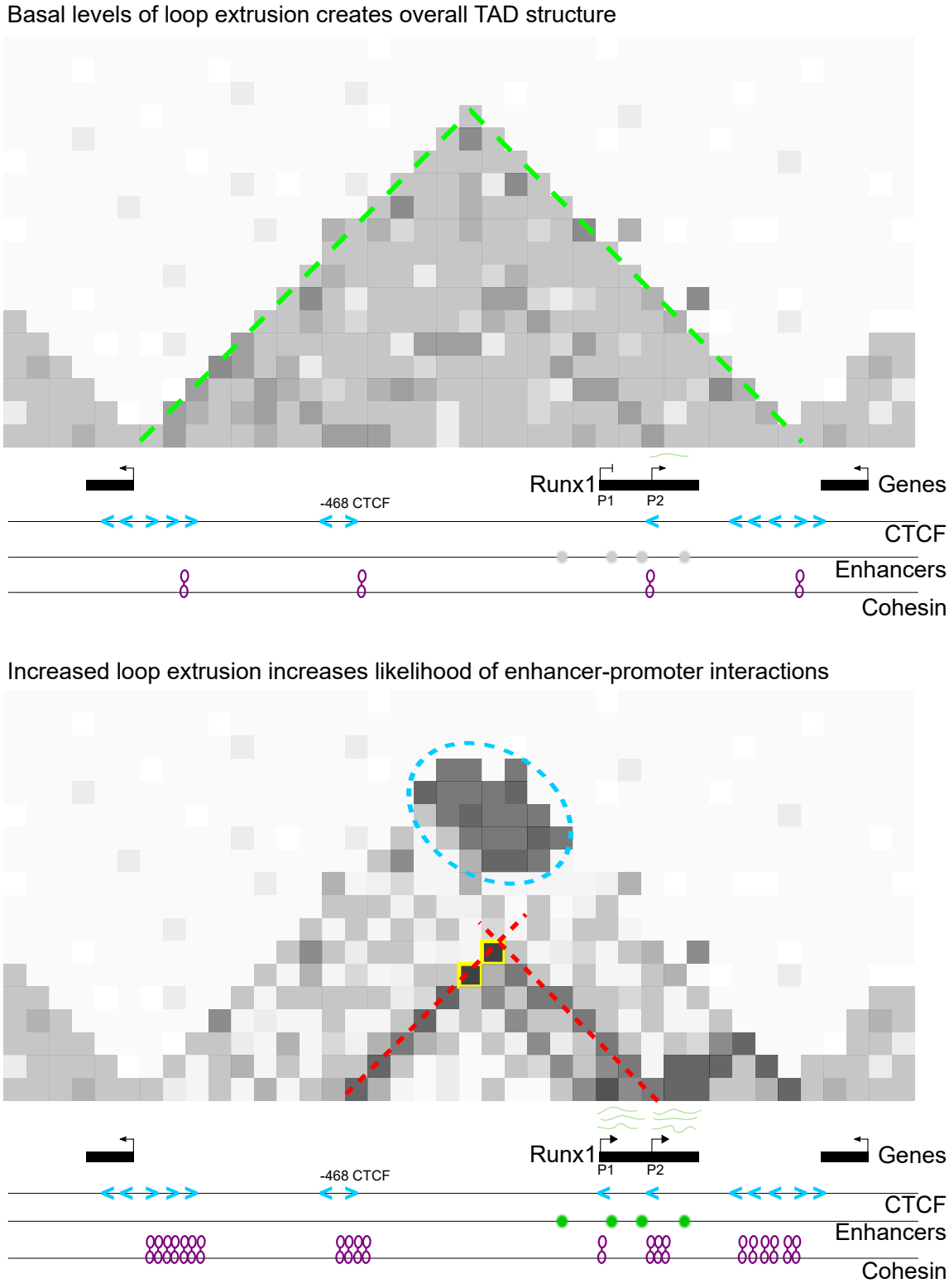


Figure 6.1 – Model of *Runx1* upregulation mediated by increased loop extrusion. *Legend continued on next page.*

Figure 6.1 – *Legend continued from previous page.* Schematic *cis*-interaction matrices are shown to represent imagined all-vs-all Hi-C or Tiled-C data. These simplified diagrams are based on real Tiled-C data at *Runx1* in E14 mESCs and 416B cells. The position of the *Runx1* gene is indicated with two neighbouring genes outside of its regulatory domain. The green wavy lines next to the promoters represent the levels of transcription from each promoter. CTCF binding sites are indicated by blue arrowheads with the orientation of the motif indicated by the direction of the arrowhead. Inactive enhancers are indicated by grey circles and active enhancers are indicated by green circles. Cohesin complexes are indicated as purple double rings. In the top panel, basal levels of loop extrusion creates the overall *Runx1* TAD (as outlined with green dashed lines) in undifferentiated mESCs. Runx1 expression is low and only occurring from the P2 promoter. CTCF is binding at the P2 promoter but not at the P1 promoter. In the lower panel, an increased frequency or processivity of loop extrusion is taking place. This generates 'stripes' of pervasive interactions throughout large tracts of the domain (red dashed lines) especially at the CTCF-bound P2 promoter and -468 CTCF site. A 'cloud' of interactions representing diffuse interactions between the two clusters of convergently oriented CTCF sites at the boundaries of the domain is outlined by a dashed blue circle. Discrete foci of interactions between CTCF sites are shown by yellow outlined squares.

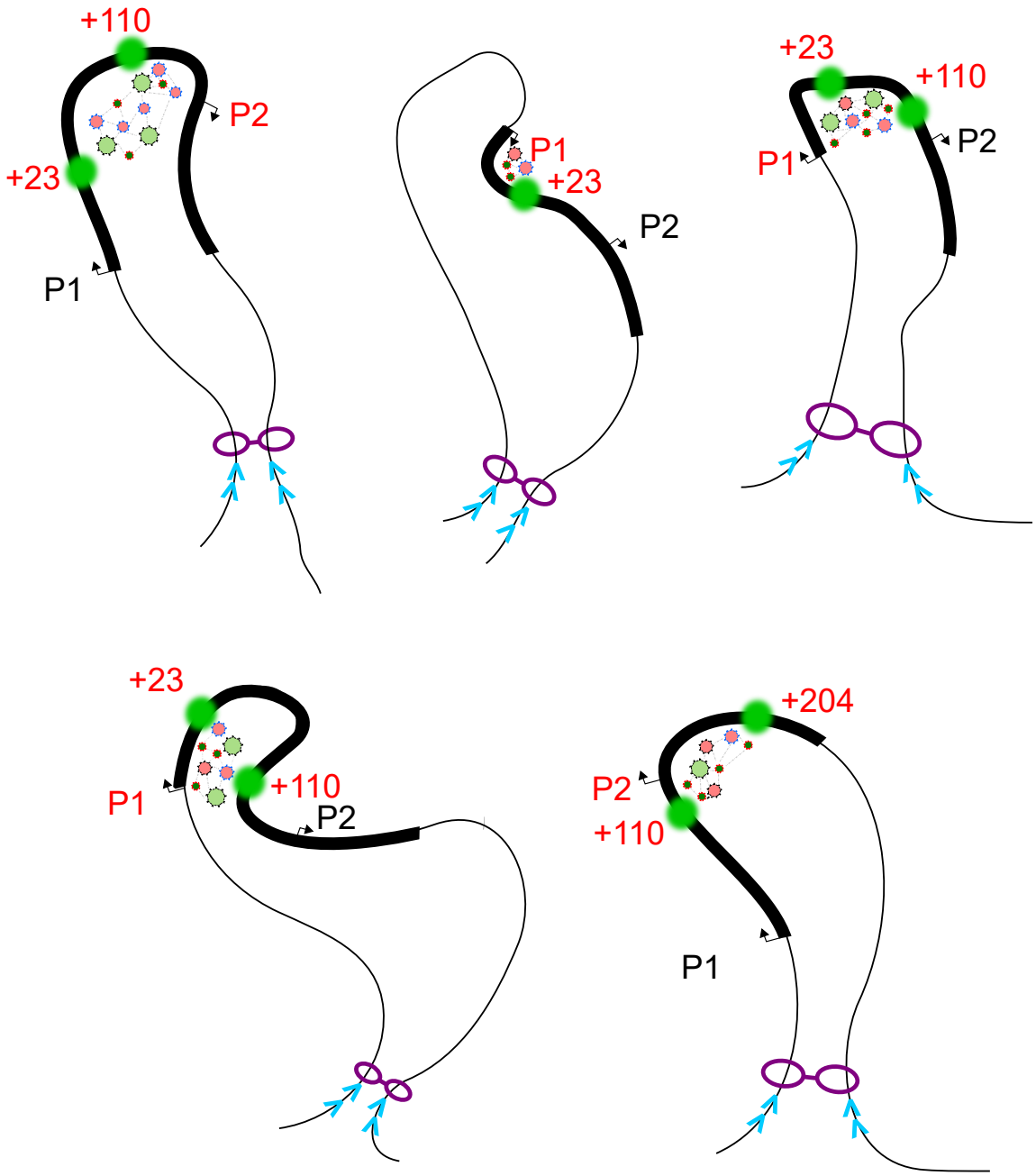


Figure 6.2 – Model of *Runx1* enhancer-promoter interactions maintained by phase separation of transcriptionally active components. *Legend continued on next page.*

Figure 6.2 – *Legend continued from previous page.* Simplified schematic of enhancer-promoter interactions stabilised by protein-protein interactions between multiple transcriptional activator complexes bound at enhancers and promoters. In each loop configuration depicted, one active enhancer is shown as a green circle labelled with the distance from the *Runx1* start codon (ATG). Both *Runx1* promoters are labelled. The thin black line represents the chromatin and the thick black line represents the *Runx1* gene. The enhancer-promoter combination in each diagram is highlighted with red text. Transcriptional activators are shown as pink green and blue circular objects between the enhancers and promoters. Grey dashed lines represent the weak multivalent interactions typical of phase separated condensates. Convergent CTCF sites are shown as blue arrowheads at the bottom of the chromatin loop. The cohesin that extruded this loop is shown as a purple double ring at the bottom of the loop. This simplified model does not capture the cohesin complexes that will be extruding loops within loops at multiple length scales continuously.

6.1.4 Unique and redundant roles for haematopoietic *Runx1* enhancers

The *Runx1* enhancers appear not to conform to a 'shadow enhancer' model. This is because they exhibited dynamic open chromatin profiles during EHT *in vitro* and were regulated by a distinct set of upstream TFs. Further, deletion of the +110 but not the +204 enhancer in mESCs led to reduced Runx1 levels in differentiated haematopoietic cells. *Runx1* enhancers appear be acting additively not synergistically, since loss of the +110 enhancer moderately impacted Runx1 levels. This is inconsistent with the enhancers forming a super enhancer (SE), as was previously suggested (Mill et al., 2019; Gunnell et al., 2016; Schuijers et al., 2018; Hnisz et al., 2017; Saint-André et al., 2016; Kwiatkowski et al., 2014). Together, it seems that the enhancers perform distinct regulatory functions during EHT. However, it is also likely the case that no single enhancer is responsible for the entirety of *Runx1* expression in any given cell, since multiple enhancers will provide phenotypic robustness (Frankel et al., 2010) and a allow greater levels of transcriptional fine-tuning. Therefore, partial enhancer redundancy is most likely the case—with each one contributing a certain fraction of the 'total *Runx1* enhancer activity' (Figure 6.3). Over EHT, the relative contributions of each of the enhancers might change (Figure 6.3), possibly due to differences in the relative concentrations of specific upstream TFs.

According to this partial redundancy model, if a single enhancer is lost then sufficient *Runx1* transcription levels could be maintained as long as a minimum critical threshold of 'total *Runx1* enhancer activity' can be provided by the other enhancers active at that stage. This model is being tested by deleting multiple enhancers in individual mESC lines. If loss of more than one enhancer leads to even greater reductions in *Runx1* transcription levels, then this will provide evidence for partial redundancy between the enhancers. Loss of +23 enhancer alongside +204 (which were both active in HE cells) would be expected to reduce *Runx1* levels primarily in HE cells. Alternatively, loss of +23 and +110 (which were both active in HP cells) would be expected to reduce expression primarily in HP cells where these enhancers were both active. However, interpreting multiple enhancer deletions performed in the same line will require careful interpretation. Even if a cell is suffering from reduced *Runx1* transcription at a particular stage of EHT when a missing enhancer would normally be dominant, accumulation of transcripts could still occur which might allow the cell to progress to the next stage where another enhancer may become more dominant. This compensation by transcript accumulation would likely only be possible for cellular intermediates like HE, and is unlikely to help more mature HP cells that might require higher levels of *Runx1* expression for extended periods of time. Indeed, this could provide an explanation for why +204 enhancer deletion showed little effect on *in vitro* EHT. It will also be valuable to assess enhancer requirements in a more defined developmental context than *in vitro* systems such as in cellular intermediates of EHT *in vivo*. Enhancer-deletion mouse lines could be generated to examine this.

6.1.5 Potential therapeutic applications of manipulating *Runx1* enhancer activity

An increased understanding of how multiple enhancers orchestrate *Runx1* expression across haematopoietic development could be useful for translational research in several

different ways. Firstly, understanding the signals converging on *Runx1* enhancers during EHT could inform directed differentiation protocols for the *de novo* generation of HSCs *in vitro*. For example, Klf/SP1 motifs were important for +23 enhancer activity and AP-1 motifs were required for *Runx1* +204 enhancer activity, and both of these enhancers showed activity in HE cells *in vitro*. Future *in vitro* experiments could apply this knowledge by using small molecule inducers of Klf factor expression and activators of AP-1 signaling to upregulate endogenous *RUNX1* expression. This could be achieved using small molecule activators of protein kinase A signaling (Lo et al. 2012, which activates AP-1 signaling, Karin 1995), and activators of AMP-activated protein kinase (which was upstream of *Klf2* expression in endothelial cells Wang et al. 2006). These small molecules could be used to replace exogenous *RUNX1* transgenes in existing differentiation cultures for the generation of HSCs (Sugimura et al., 2017). Whether targeting these pathways induces endogenous *RUNX1* expression, promotes EHT, and facilitates the generation of transplantable HSCs should be investigated.

Secondly, *RUNX1* mutations and chromosomal translocations are associated with a wide variety of human leukaemias (Takei and Kobayashi, 2019). The *Runx1* +23 and +110 enhancers were specifically active in HP cells during mouse developmental haematopoiesis and in 416B myeloid progenitor cells, suggesting that the human +23 and +110 enhancers may also be driving *RUNX1* expression in leukaemia. Therefore, strategies to reduce *RUNX1* expression could be developed by targeting these enhancers using CRISPR/Cas9. Further, the fact that Klf and Cebp motifs were important for the activities of these enhancers could be exploited by targeting only these TFBS specifically. A precise editing approach could be applied to generate a single DNA double strand break (DSB) targeted to a specific TFBS within the enhancers. This could help reduce the occurrence of Cas9-induced LDs, as multiple adjacent DSBs was associated with increased LD formation compared to a single DSB (Owens et al., 2019). Together, a greater understanding of *Runx1* transcriptional regulation could inform protocols for the *de novo* generation of HSCs *in vitro* and refine strategies for targeting *RUNX1* in human leukaemia.

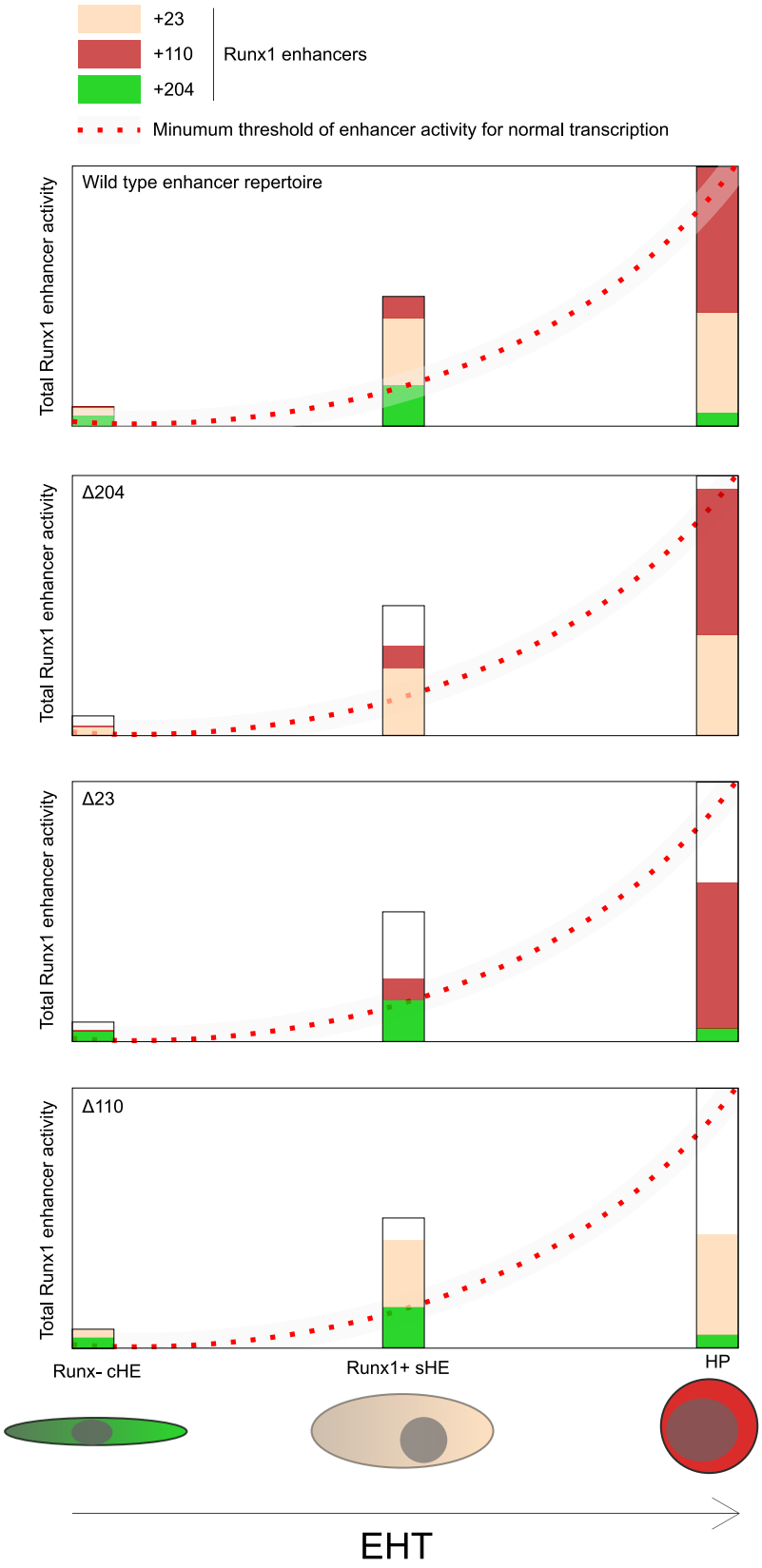


Figure 6.3 – Model of partial enhancer redundancy in the *Runx1* locus. *Legend continued on next page.*

Figure 6.3 – *Legend continued from previous page.* Each set of bars represents the 'total *Runx1* enhancer activity' at a given stage of cellular development during endothelial-to-haematopoietic transition (EHT) as estimated from the known increase in *Runx1* expression levels during this time, and increase in chromatin accessibility and TF binding at *Runx1* enhancers. As such, the sum total of *Runx1* enhancer activity increases during EHT, as reflected by the increased height of the bars. The relative contributions of each enhancer to overall *Runx1* enhancer activity changes during EHT. In this model, at the competent haemogenic endothelium (cHE, just before *Runx1* transcription begins), the main enhancer becoming active is +204 (green bars). This was inferred from the fact that the +204 enhancer shows an open chromatin peak in HE cells and drives enhancer-reporter expression only in HE cells (de Bruijn lab, unpublished observations). At Runx1+ specifying HE (sHE) the dominant enhancer is +23 (beige bars), as implied by the fact that the +23 enhancer becomes a strong open chromatin peak in HE cells and drives enhancer-reporter expression in HE cells and in all populations undergoing EHT *in vivo* (Nottingham et al., 2007; Bee et al., 2009b; Swiers et al., 2013a). In Runx1+ CD41+ HP cells, the +110 enhancer is proposed to be dominant. This is inferred from the fact that +110 drives enhancer-reporter expression exclusively to HP cells, and upstream regulators associated with HP cells (such as Cebp factors) were shown to be crucial for enhancer function. The minimum required threshold for *Runx1* enhancer activity is imagined to increase over time as *Runx1* transcription levels increase, and is illustrated as the red dashed line. The grey margin illustrates the potential ability for transcript accumulation to overcome a mild deficit in enhancer activity.

6.1.6 Regulation of *Runx1* alternative promoter choice

It remains to be determined what drives the P2 to P1 promoter switch observed during developmental haematopoiesis (Pozner et al., 2007; Sroczynska et al., 2009a; Bee et al., 2009b, 2010). One possibility is that antagonism exists between the two promoters, mediated by transcription-induced DNA methylation changes and CTCF-binding modulation. In such a promoter antagonism model, the P2 promoter is unmethylated in pluripotent cells possibly in part due to low levels of transcription from P2 (it is a bivalent promoter in these cells, [1] Figure 6.4). The CTCF binding site upstream of the P2 promoter will be unmethylated and bound by CTCF, which might facilitate enhancer-promoter interactions and promoter activity (Schuijers et al., 2018). During haematopoietic differentiation, *Runx1* expression in HE cells is initiated from the P2 promoter by upstream signals activating *Runx1* enhancers. At this stage, the P1 remains methylated, unbound by CTCF, and transcriptionally silent ([2] Figure 6.4). Then, demethylation of the P1 could occur via one of several alternative scenarios (as discussed in section 5.3), resulting in binding of CTCF, and transcriptional activation ([3] Figure 6.4). Ultimately the activation of P1 would lead to read-through transcription over the P2 CGI promoter which likely leads to its methylation (Jeziorska et al., 2017), inhibition of CTCF binding and silencing.

The notion that P1 promoter activity antagonises P2 activity begs the question of how the P2 and P1 promoters could ever be active at the same time, as was seen in populations of 416B cells ([4] Figure 6.4). One explanation is that single alleles within the cell population transcribe either from the P1 or P2 promoter mutually exclusively. Single cell gene expression analysis would uncover whether both promoters can be active at the same time in a single cell, but not at the level of single alleles. It seems unlikely that the P2 promoter was silent and methylated even in a subset of 416B cells because at the population level the P2 promoter was fully demethylated. An alternative explanation for how the P2 promoter may remain unmethylated in 416B cells could be the ongoing maintenance of transcription. Indeed, if sufficient levels of transcription from CGI promoters is maintained, then Dnmt3a/b may still be excluded (Baubec et al., 2015; Jeziorska et al., 2017). In this way, the two *Runx1* promoters might antagonise each other on a regulatory 'seesaw' involving DNA methylation, CTCF binding, enhancer-promoter interactions. As depicted in the final stage of the model ([5] Figure 6.4) if transcription becomes favoured at P1 compared to P2 (possibly due to changes in enhancer-promoter interactions), then the P2 promoter will no longer be able to maintain sufficient transcription levels and will become methylated, CTCF will no longer bind, and the promoter will be inactivated.

A possible causal role of CTCF sites upstream of the *Runx1* promoters in alternative promoter choice will be investigated using the mESC lines that I generated which lack these CTCF sites. Directed *in vitro* differentiation of knock-out clones will be performed to assess the effect of the deletions on haematopoiesis, alongside *Runx1* promoter-specific gene expression analysis. If the CTCF sites do play a role, then dysregulation of the promoters should be seen. Tiled-C analysis in haematopoietic cells derived from promoter CTCF knock-out clones will reveal whether the chromatin boundary observed at P2 is CTCF-dependent. If deleting the CTCF sites upstream of the *Runx1* promoters has no effect on transcription or on sub-TAD boundary formation, then this will be an interesting result because sub-TAD formation would

then depend on transcription from the promoter and not CTCF binding. Conversely, if deleting the CTCF sites does decrease promoter activity, then it will be clear that the CTCF sites do play a role in promoter activity (which may or may not also lead to altered sub-TAD boundary formation). Bisulfite sequencing of the two *Runx1* promoters will also be done in CTCF deleted mESCs (both undifferentiated mESCs and differentiated haematopoietic cells) to determine whether CTCF binding might also be a mechanism employed by the promoters to inhibit their DNA methylation. Together these experiments will help tease apart the relationship between sub-TAD boundary formation, CTCF binding, DNA methylation, and promoter activity.

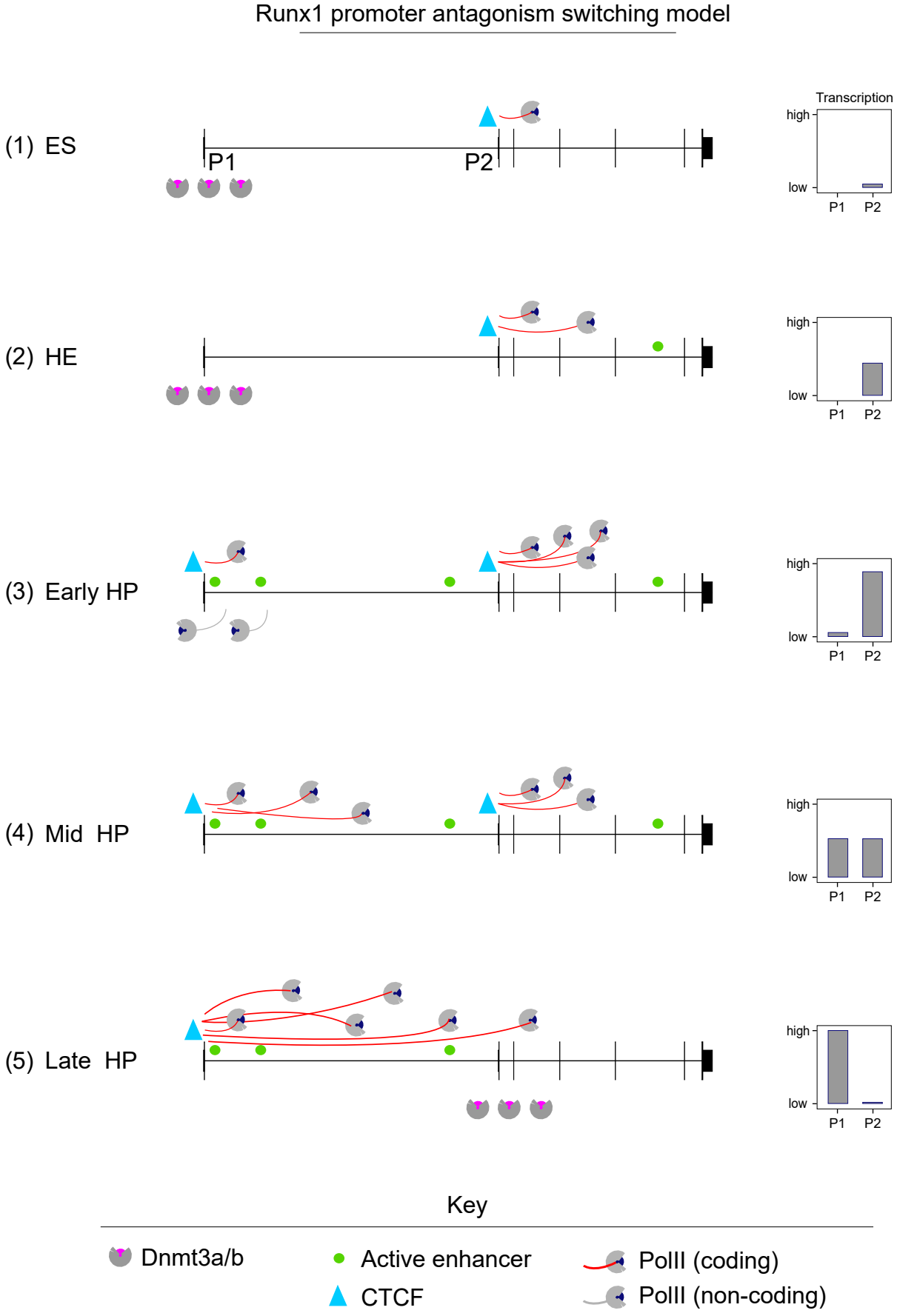


Figure 6.4 – Promoter antagonism model of *Runx1* promoter switching during embryonic development. *Legend continued on next page.*

Figure 6.4 – *Legend continued from previous page.* A model of how the observed developmentally regulated switch from P2 derived to P1 derived *Runx1* transcription (Pozner et al., 2007; Sroczynska et al., 2009a; Bee et al., 2009b, 2010) might occur. The two promoters of the gene are labelled with CTCF sites upstream of the promtoers shown as blue triangles. Active enhancers are shown as green circles. DNA methyla-transferases Dnmt3a/b are shown as dark grey and pink inverted pac men to indicate regions that are methylated. Transcribing RNA polymerase II complexes are shown as light grey and dark blue pac men with red tails to indicate coding transcription coming from the promoters or with grey tails to indicate anti-sense non-coding transcription coming from the enhancers. The relatively low level of P2 derived transcription in un-differentiated mESCs was determined experimentally by RNA-seq. The relative level of transcription in the different haematopoietic lineages are inspired by the known relative levels in early to late developmental haematopoiesis (Pozner et al., 2007; Sroczynska et al., 2009a; Bee et al., 2009b, 2010).

7. Appendix

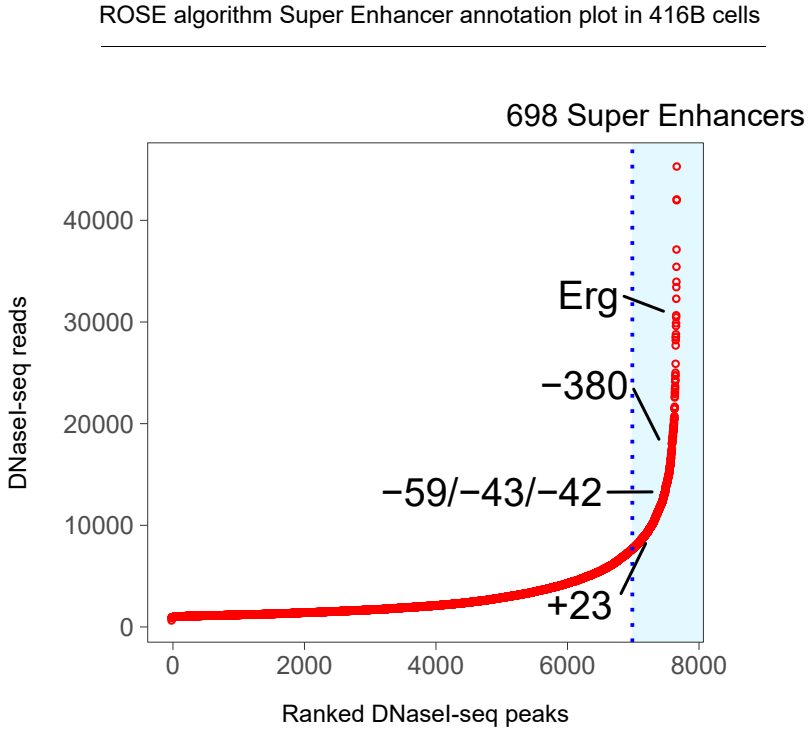


Figure 7.1 – Calling super enhancers in 416B cells. DNaseI peaks were called in 416B cells and used alongside 416B DNaseI-seq read depth to call the top super enhancers (SEs) in 416B cells using the ROSE algorithm (Loven et al., 2013; Whyte et al., 2013). Labels indicate the *Runx1* enhancers that were classed as SEs and an alternative promoter region of *Erg*. Each red dot is a cluster of stitched DNaseI-seq peaks (see methods section 2.9.5) and the dashed vertical line represents the cut-off used for calling SEs.

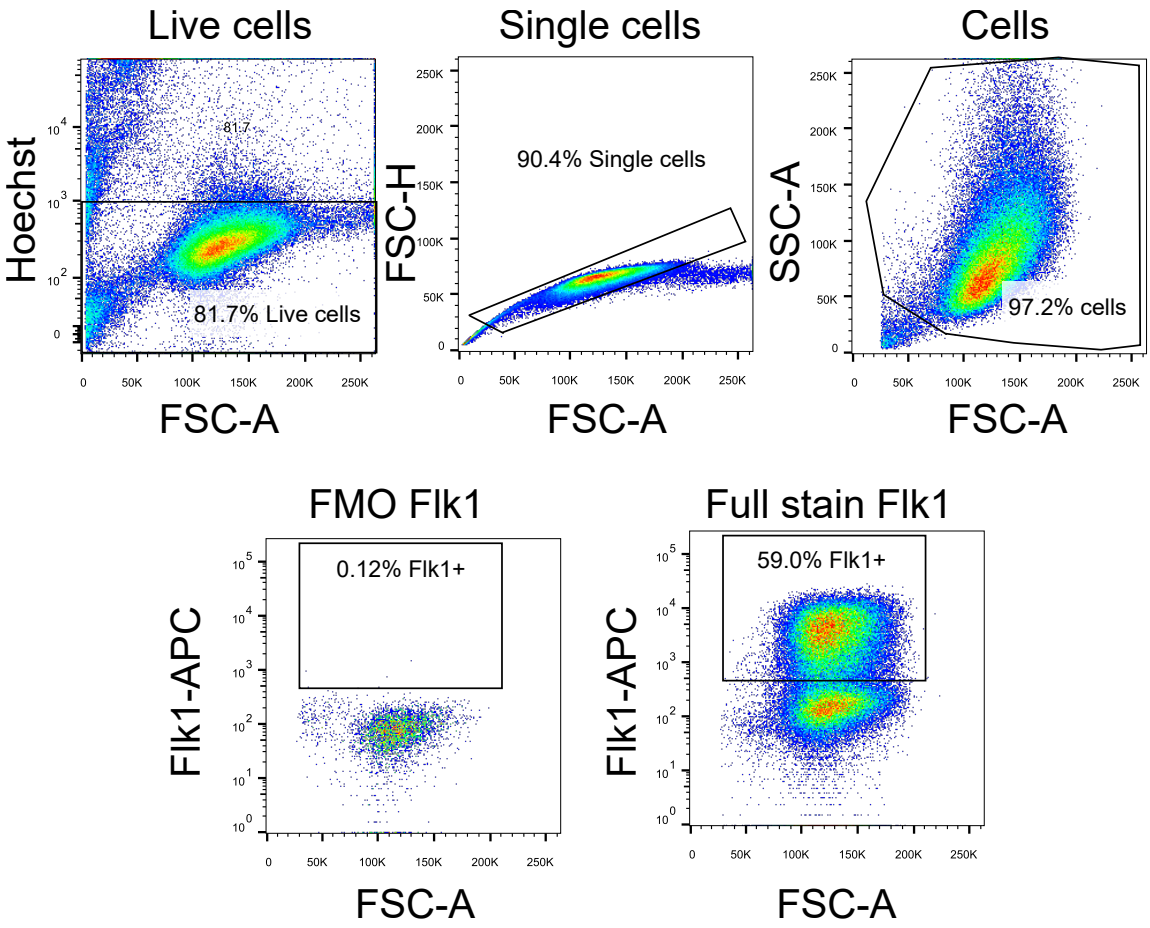


Figure 7.2 – Analysis and representative FACS plots showing gating strategy to isolate Flk1+ cells. Staining was done in day 4 mESCs differentiated to mEBs and disaggregated. Full stained cells were stained with Hoechst and Flk1-APC and FMO Flk1 cells were stained with Hoechst. Sorted cells were gated on live single Flk1+ cells. One representative differentiation is shown. Flk1+% varied from 20-80% between individual differentiation cultures.

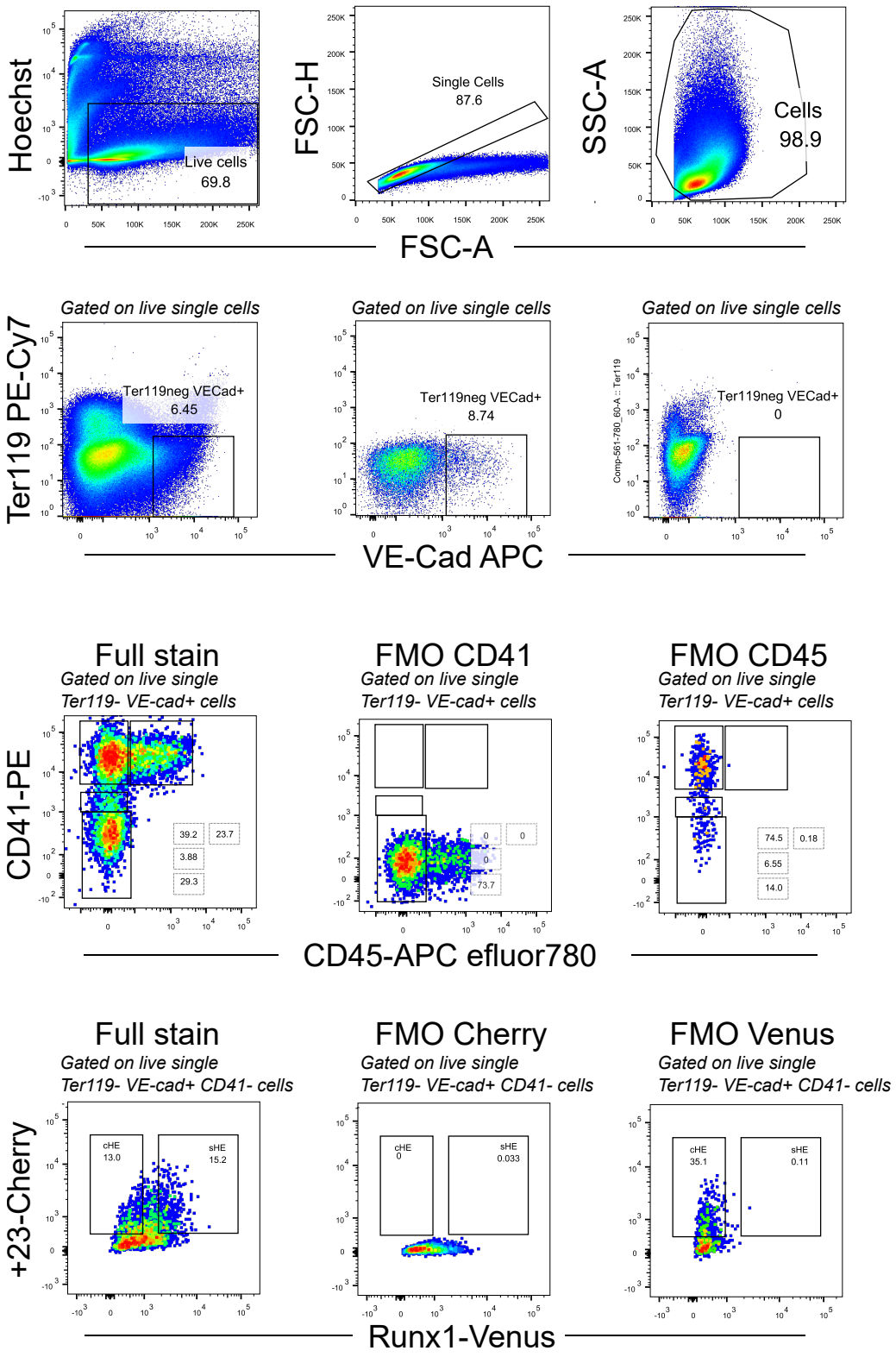


Figure 7.3 – Analysis and representative FACS plots of immunostaining in day 7 differentiated mESCs. *Legend continued on next page.*

Figure 7.3 – *Legend continued from previous page.* Sorted cHE cells were gated on live single Ter119- VE-cad+ CD41- 23Cherry+ Runx1Venus- cells. Sorted sHE cells were gated on live single Ter119- VE-cad+ CD41- 23Cherry+ Runx1Venus+ cells. Sorted HP cells were gated on Ter119- VE-cad+ CD41+. One representative differentiation is shown.

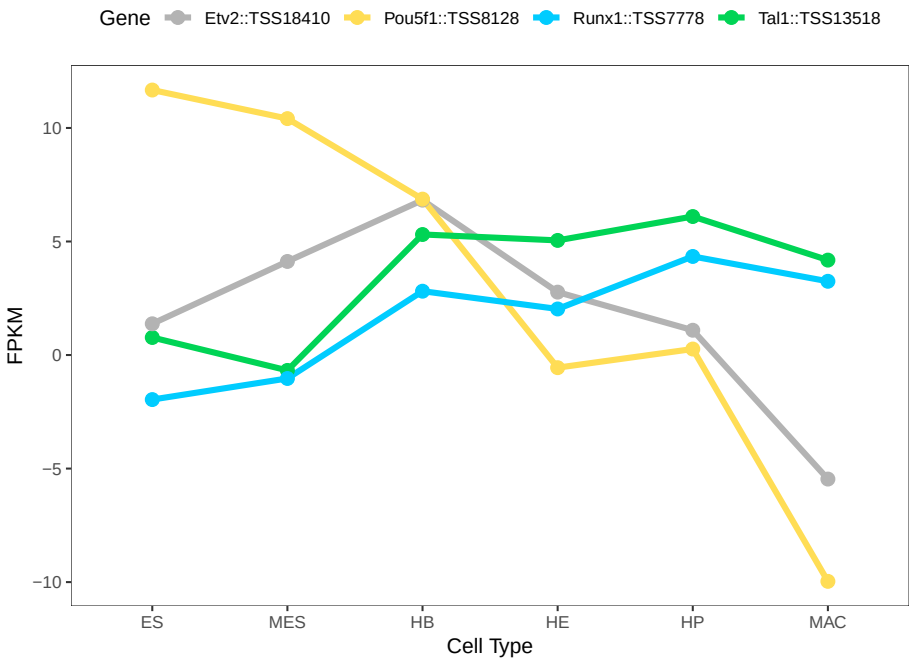


Figure 7.4 – RNA-Seq expression of *Runx1* during *in vitro* endothelial to haematopoietic transition. Data were previously generated by Goode et al. 2016. The expression of pluripotency marker *Pou5f1* and endothelial marker *Etv2* both decreased during differentiation, while the expression of haematopoietic genes *Runx1* and *Tal1* both increased.

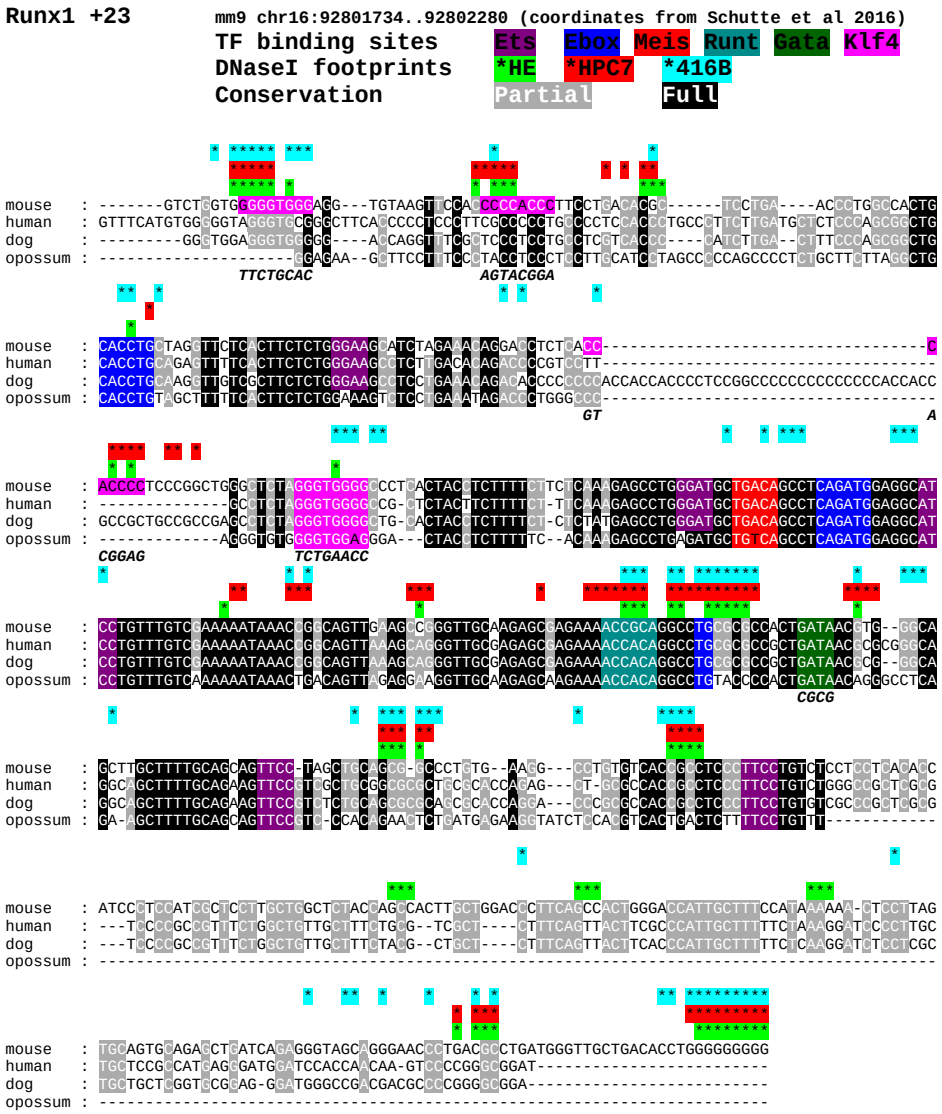


Figure 7.5 – Full +23 enhancer sequence alignment as in Schütte et al. (2016) showing mutations done to binding sites. *Legend continued on next page.*

Figure 7.5 – *Legend continued from previous page.* TF binding sites previously identified are shown (Ets, Ebox, Meis, Runt, Gata Schütte et al. 2016). Four newly identified Klf/Sp1 binding motifs are highlighted in pink. A highlighted star above each base indicates that a DNaseI footprint was identified at that position. Green highlighted stars were digital DNaseI footprints in HE cells (Goode et al., 2016), red highlighted stars were identified in HPC7 cells (Wilson et al., 2010a), and blue highlighted stars were footprints in 416B cells (Vierstra et al., 2014). Mutations performed to Klf and Gata sites (control mutation known to abrogate enhancer activity) are indicated underneath the sequence alignments.

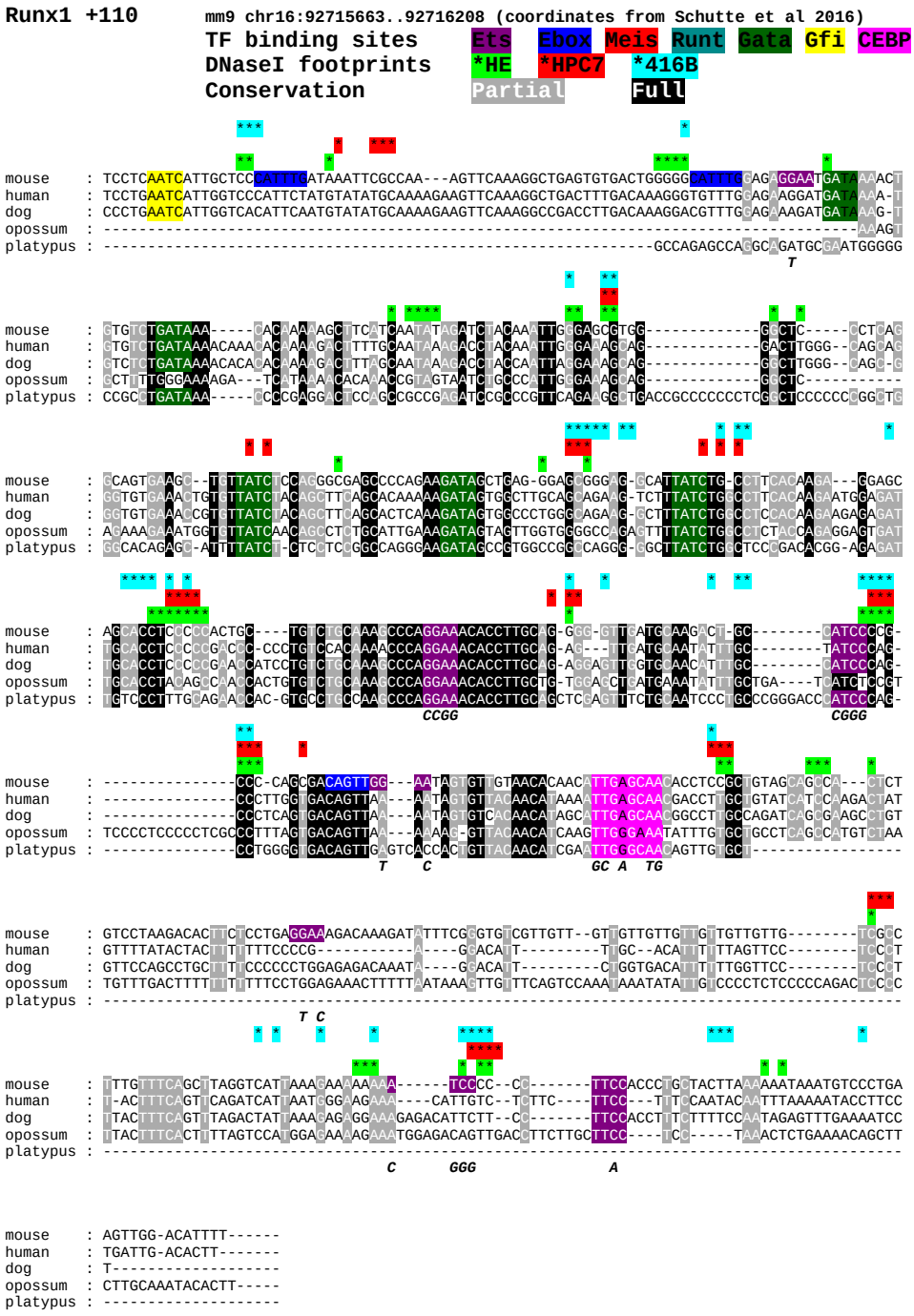


Figure 7.6 – Full +110 enhancer sequence alignment as in Schütte et al. (2016) showing mutations done to binding sites. *Legend continued on next page.*

Figure 7.6 – *Legend continued from previous page.* TF binding sites previously identified are shown (Ets, Ebox, Meis, Runt, Gata, Gfi Schütte et al. 2016). One newly identified Cebp binding motif is highlighted in pink. A highlighted star above each base indicates that a DNaseI footprint was identified at that position. Green highlighted stars were footprints in HE cells (Goode et al., 2016), red highlighted stars were identified in HPC7 cells (Wilson et al., 2010a), and blue highlighted stars were footprints in 416B cells (Vierstra et al., 2014). Mutations performed to Cebp and Ets sites (control mutation known to abrogate enhancer activity) are indicated underneath the sequence alignments.

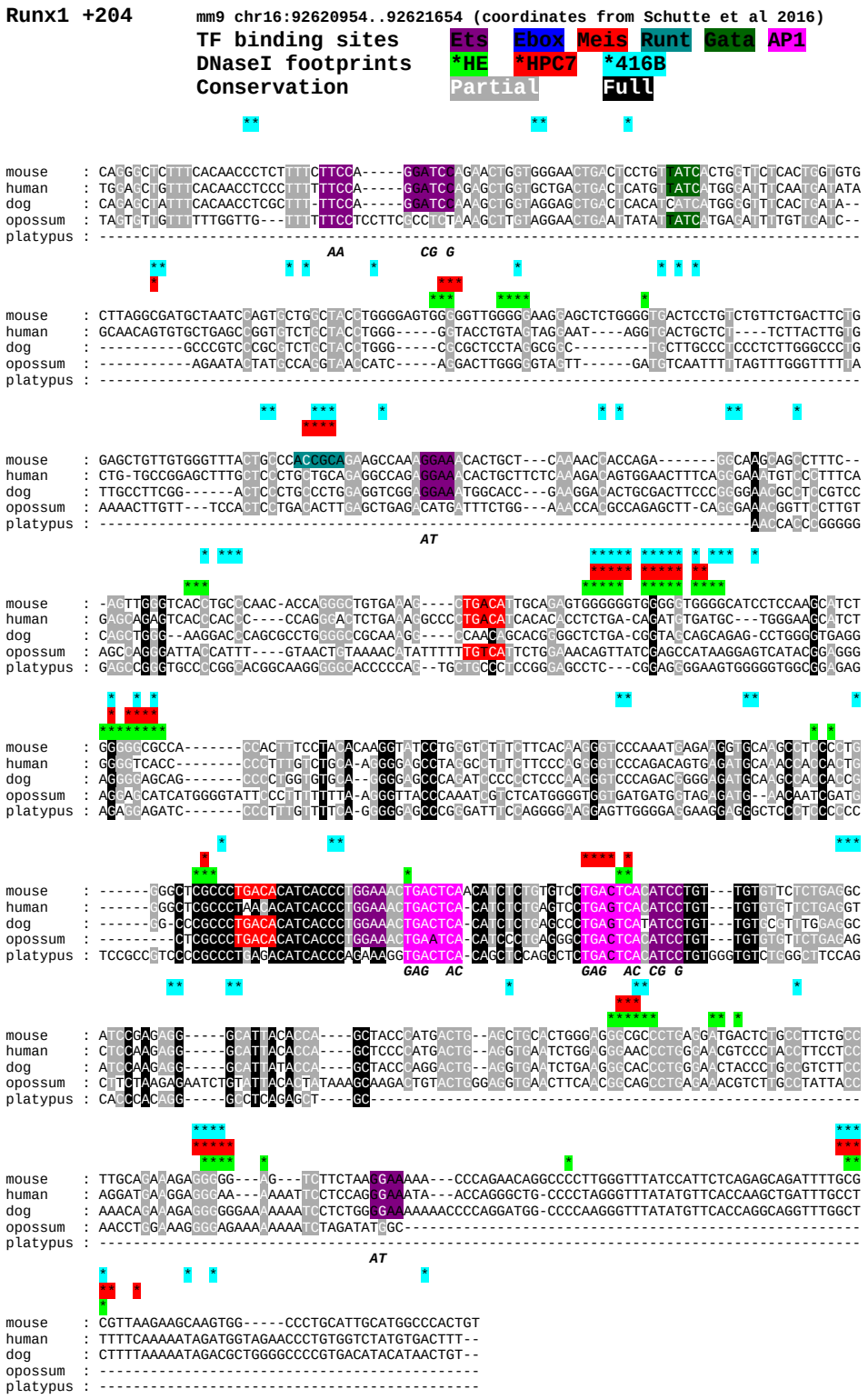


Figure 7.7 – Full +204 enhancer sequence alignment as in Schütte et al. (2016) showing mutations done to binding sites. *Legend continued on next page.*

Figure 7.7 – *Legend continued from previous page.* TF binding sites previously identified are shown (Ets, Ebox, Meis, Runt, Gata Schütte et al. 2016). Two newly identified AP1 binding motifs are highlighted in pink. A highlighted star above each base indicates that a DNaseI footprint was identified at that position. Green highlighted stars were footprints in HE cells (Goode et al., 2016), red highlighted stars were identified in HPC7 cells (Wilson et al., 2010a), and blue highlighted stars were footprints in 416B cells (Vierstra et al., 2014). Mutations performed to AP1 and Ets sites (control mutation known to abrogate enhancer activity) are indicated underneath the sequence alignments.

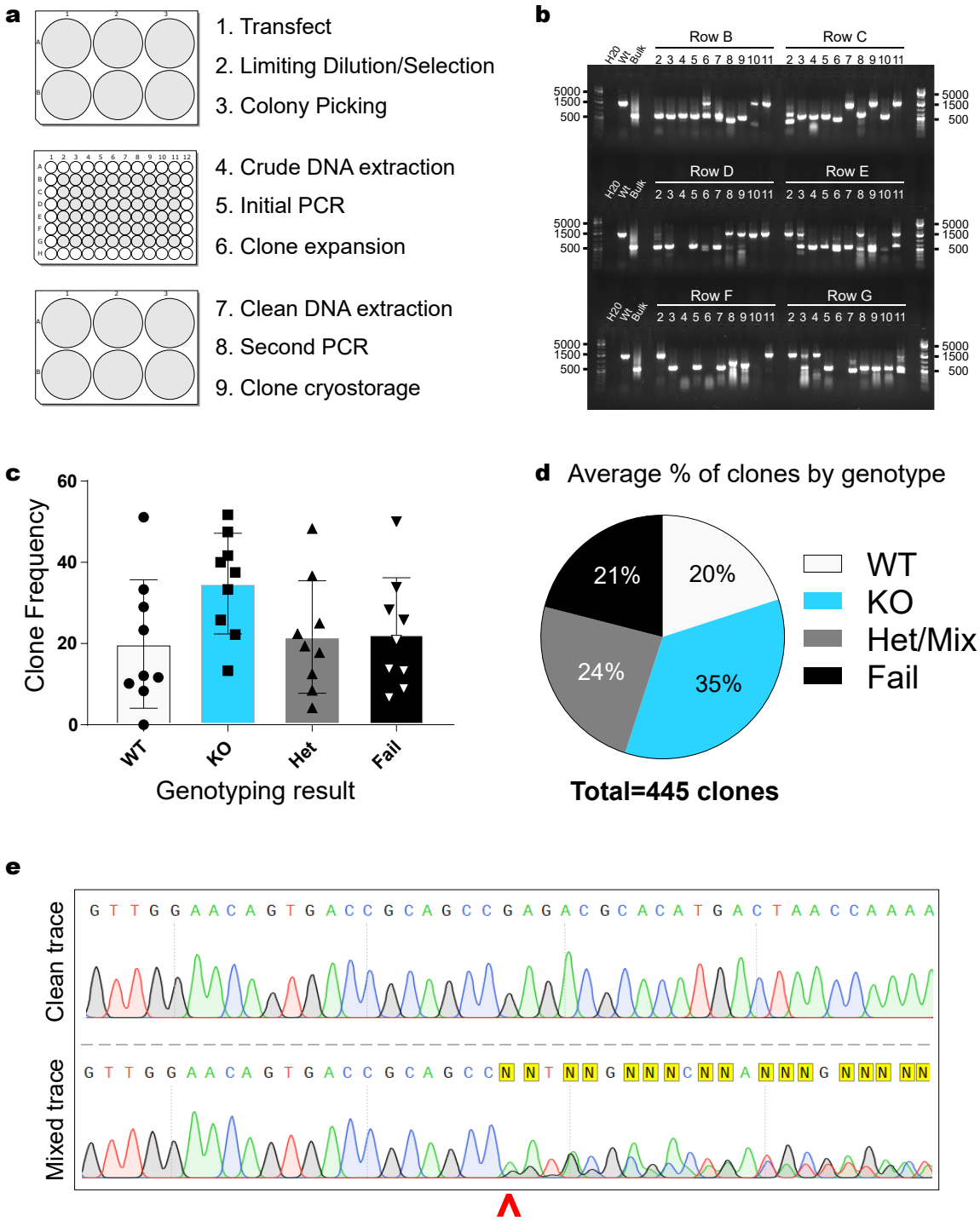


Figure 7.8 – Overview of genome editing to delete *Runx1* enhancers. *Legend continued on next page.*

Figure 7.8 – *Legend continued from previous page.* a) Protocol schematic for mESC gene-editing, colony-picking, genotyping, and expansion. b) An example short-range PCR genotyping gel result from one experiment targeting +204 in mESCs with $Cas9^{D10A}$ nickase. The wildtype amplicon size is 2 kb, and the knock-out amplicon size is 500 bp. c) Genotype frequencies from three independent editing experiments at +23, +110, +204 enhancers determined by short-range PCR screening. d) Average genotyping results from all of the three experiments at each enhancer. e) Example Sanger sequencing traces from PCR products amplified from clones that were identified as homozygous knock-out clones with short-range PCR. A typical 'clean trace' read (top trace) and 'mixed trace' read (bottom trace) are shown for comparison. The site of the beginning of the 'mixed trace' is indicated with a red arrow head.

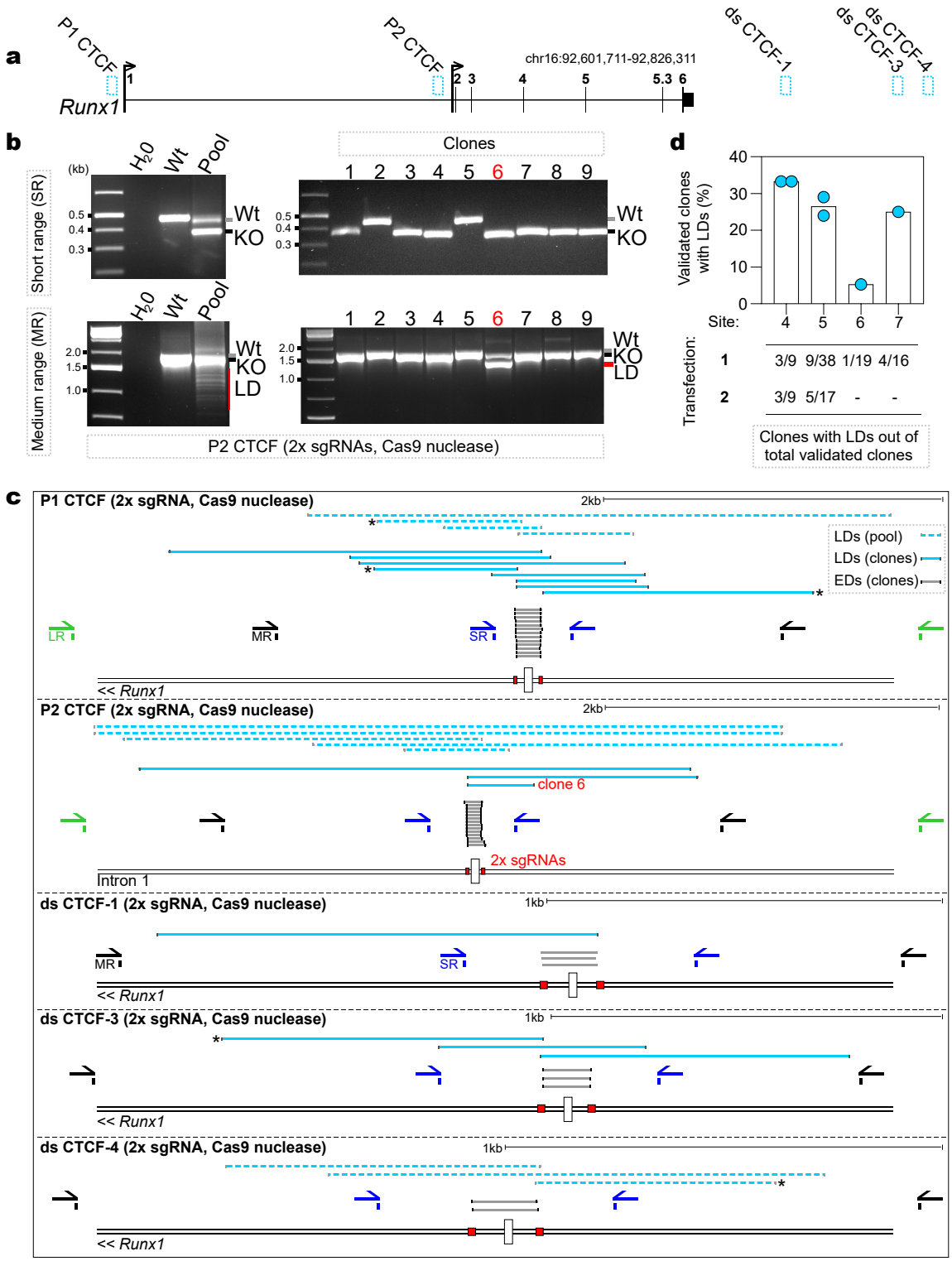


Figure 7.9 – Larger deletions are generated after gene-editing using Cas9 nuclease. *Legend continued on next page.*

Figure 7.9 – *Legend continued from previous page.* a) Locus maps of CRISPR/Cas9 nuclease strategies to delete CTCF sites in the Runx1 locus. b) Gel images showing PCR amplification of gDNA isolated from a pool of selected cells (left-hand gels) or isolated clones (right-hand gels) targeted with Cas9 nuclease at Site 5. PCR screening was performed with short-range primers (top gels) and medium-range primers (bottom gels). Wt next to the gel image indicates the size of the wild type allele, KO indicates the size of alleles harbouring the expected deletion, and LD indicates the size of alleles identified harbouring larger deletions. c) Schematic showing the positions of short-range PCR primers (SR blue), medium-range PCR primers (MR, black), longer-range PCR primers (LR, green), sgRNAs (red boxes), and LDs isolated from clones (dark blue lines) and pools of cells (light blue dashed lines). (D) Quantification of clone frequencies with homozygous wild type or knock-out genotypes by SR PCR (validated clones) that contained a LD on one allele only detected by medium-range or longer-range PCR (n=1-2 independent transfections per site, each dot is one independent experiment).

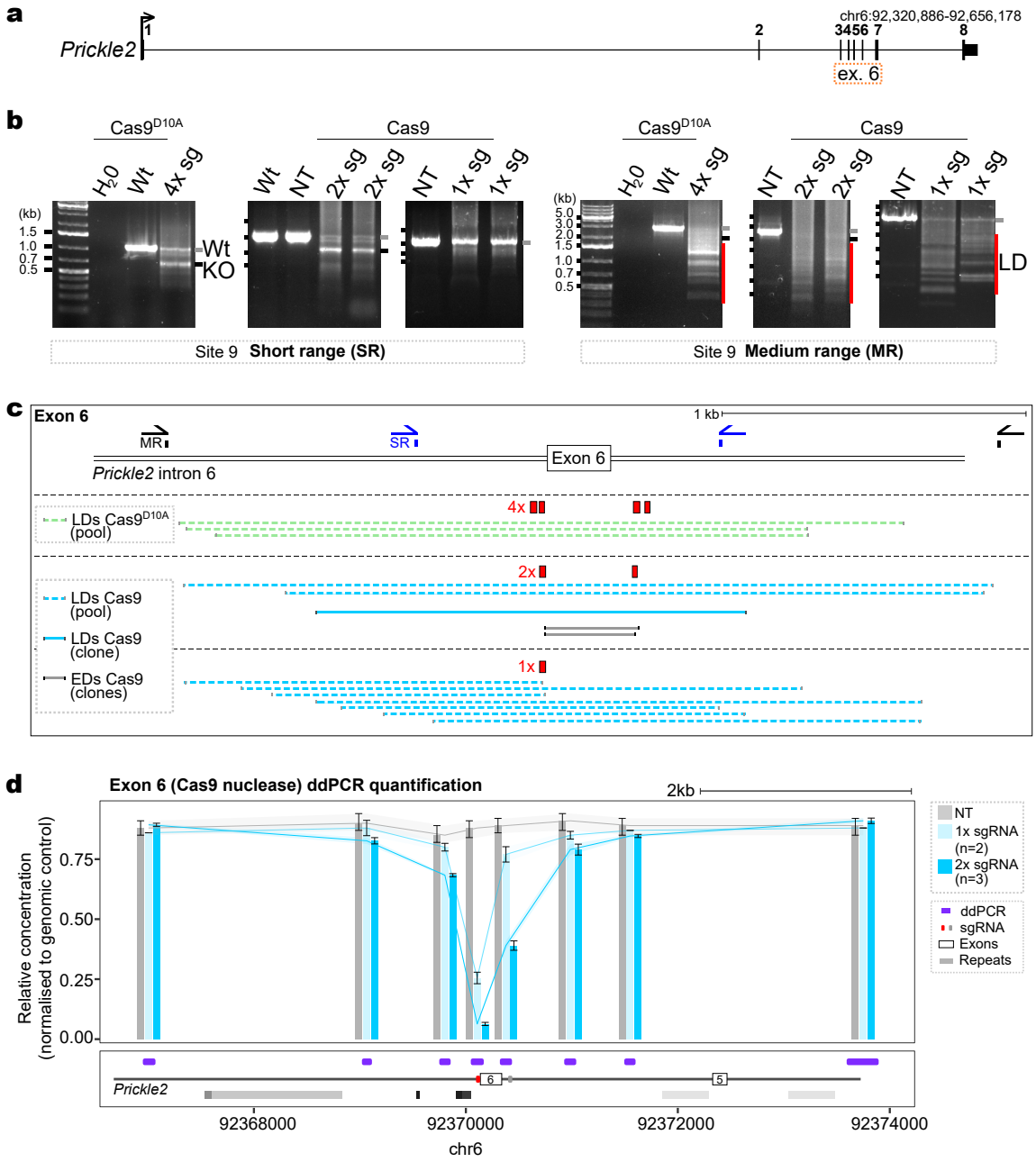


Figure 7.10 – Larger deletions are generated in a variety of genome-editing contexts.
Legend continued on next page.

Figure 7.10 – *Legend continued from previous page.* a) Locus schematic showing *Prickle2* exon 6 on mouse chromosome 6. b) Gel images showing PCR amplification products from gDNA harvested from a pool of transfected cells targeted using the CRISPR/Cas9 strategies indicated. Left gel images correspond to short-range primers and right gel images correspond to medium-range primers. Wt and a grey line next to the gel image indicates the size of the wild type allele, KO and a black line indicates the size of alleles harbouring the expected deletion (based on the location of 2x or 4x sgRNAs), and LD and a red line indicates the size of alleles identified harbouring LDs. c) Schematic showing the 4x sgRNA Cas9^{D10A} nickase, 1x and 2x sgRNA Cas9 nuclease strategies targeting *Prickle2* exon 6. Sequenced PCR products amplified from pools of cells (light blue and light green dashed lines) and one isolated cone (dark blue line). Mapped deletions of expected size (EDs) based on the location of the 2x sgRNA cut sites are shown (grey lines). d) ddPCR quantification of deletions targeting exon 6 with Cas9 nuclease and 1x sgRNA (red box) or 2x sgRNAs (red and grey boxes). Each bar represents the mean +/- 95% confidence interval. mESCs were targeted with 1x sgRNA (light green bars, n=2), 2x sgRNA (dark green bars, n=3) and non-targeting control (grey bar). Alasdair Allan performed ddPCR on genomic DNA prepared by me. Joe Harman designed *Prickle2* exon 6 deletion strategy including PCR primers, and performed some transfections.

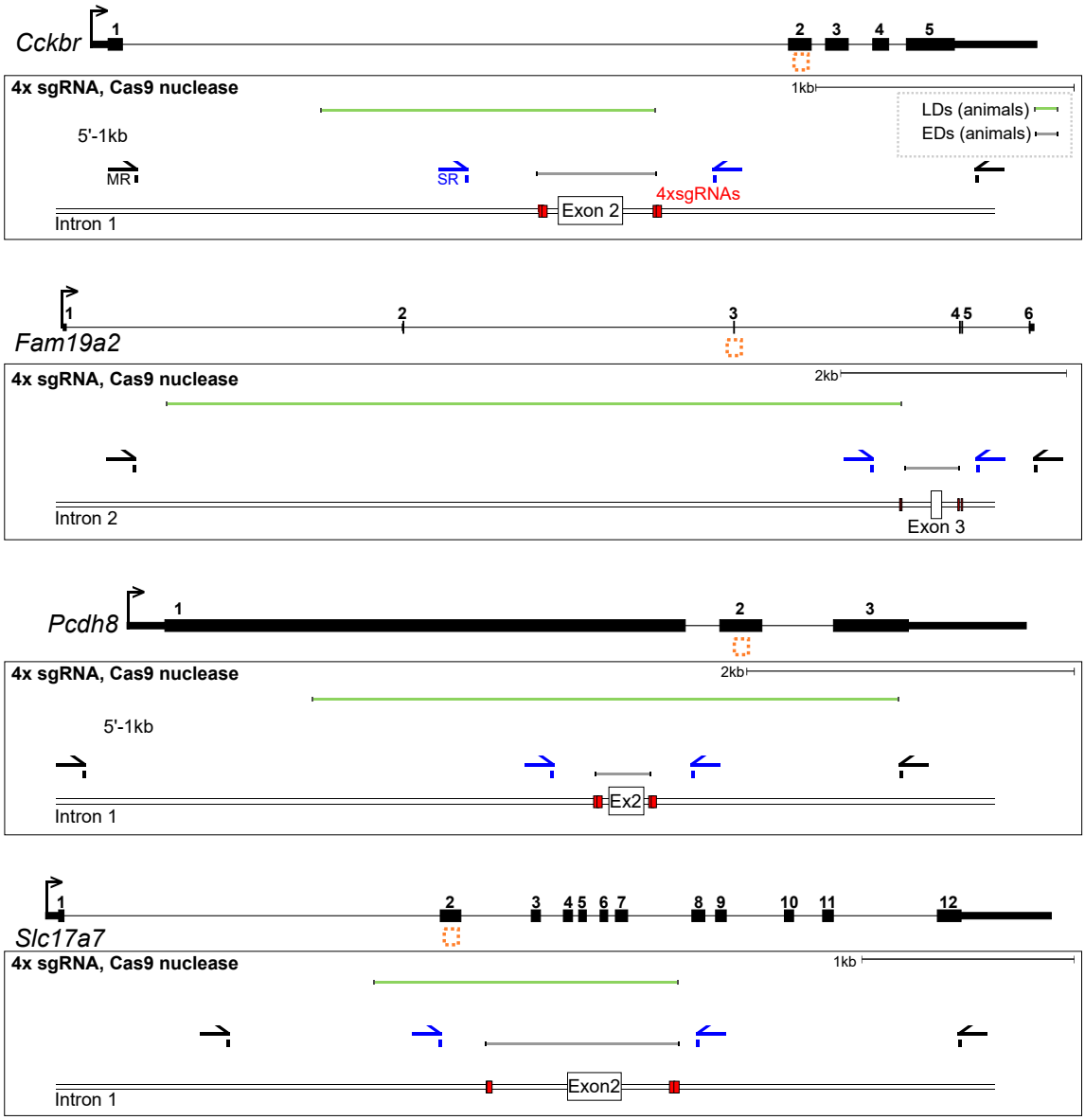


Figure 7.11 – Larger deletions when gene editing *in vivo*. Locus maps of CRISPR/Cas9 strategies to delete the genes *Cckbr*, *Fam19a2*, *Pcdh8*, and *Slc17a7*. Schematics show the positions of short-range PCR primers (SR, blue), medium-range PCR primers (MR, black), sgRNAs (red boxes), ddPCR amplicons (purple lines) and larger deletions (green lines). Deletion and screening strategies and experiments were performed by Lydia Teboul, Adam Caulder, Alasdair Allan, and Gemma Codner.

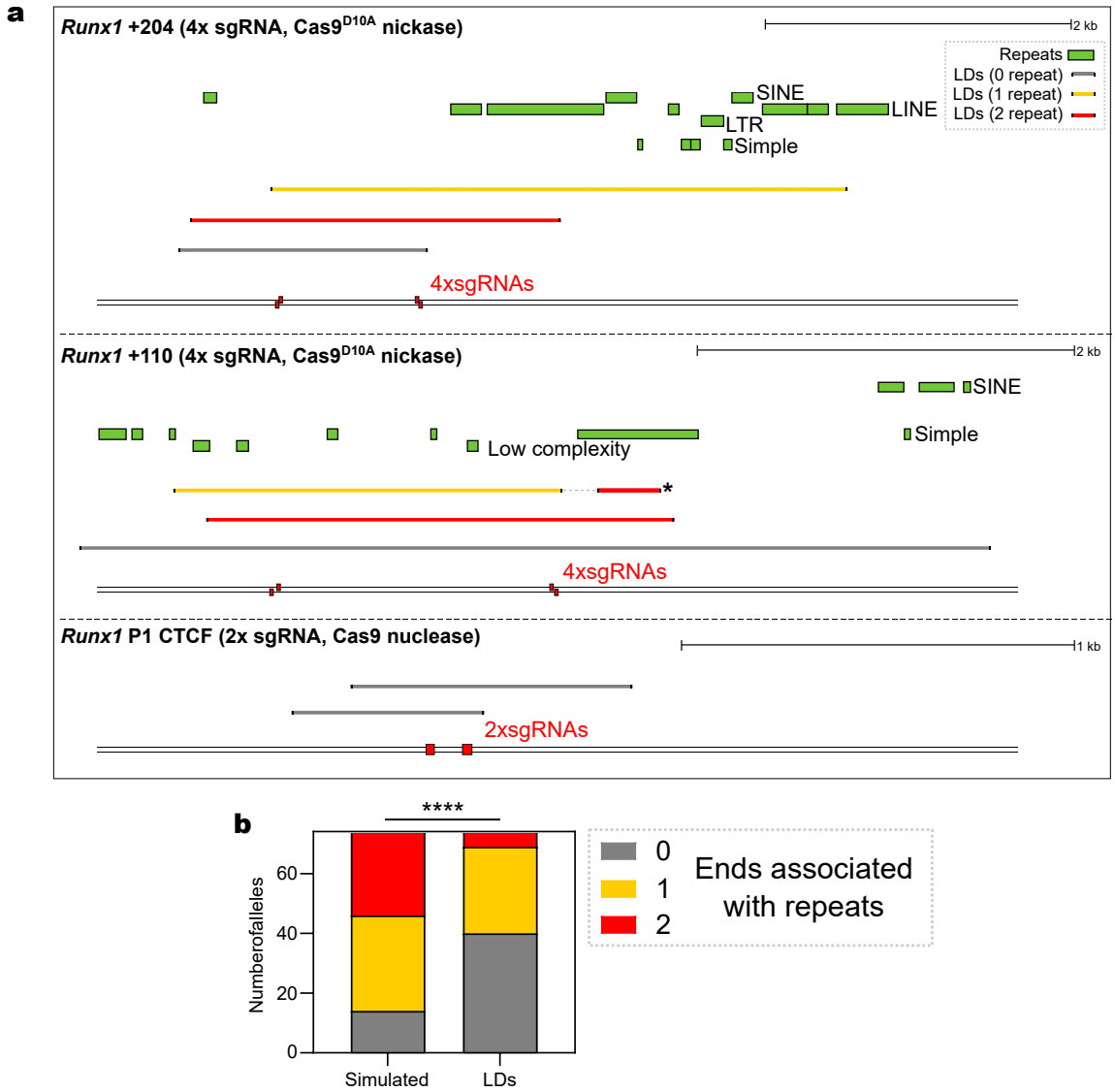


Figure 7.12 – Quantification of repeat elements associated with larger deletions. a) Annotated repetitive elements (green boxes with black outlines) were mapped alongside LDs using the UCSC genome browser RepeatMasker. LDs were defined as having neither end (grey bars), one end (orange bars), or two ends (red bars), intersecting within 100 bp of annotated repeat elements. A secondary deletion that was upstream from the original cut site and removed one of the S-R PCR primer binding sites in a clone targeted at *Runx1* +110 enhancer was contained within a simple tandem repeat (red bar marked with *) b) Quantification of repeat intersections with 74 LDs and the same number of equally sized simulated deletions (****, χ^2 test, $p < 0.0001$).

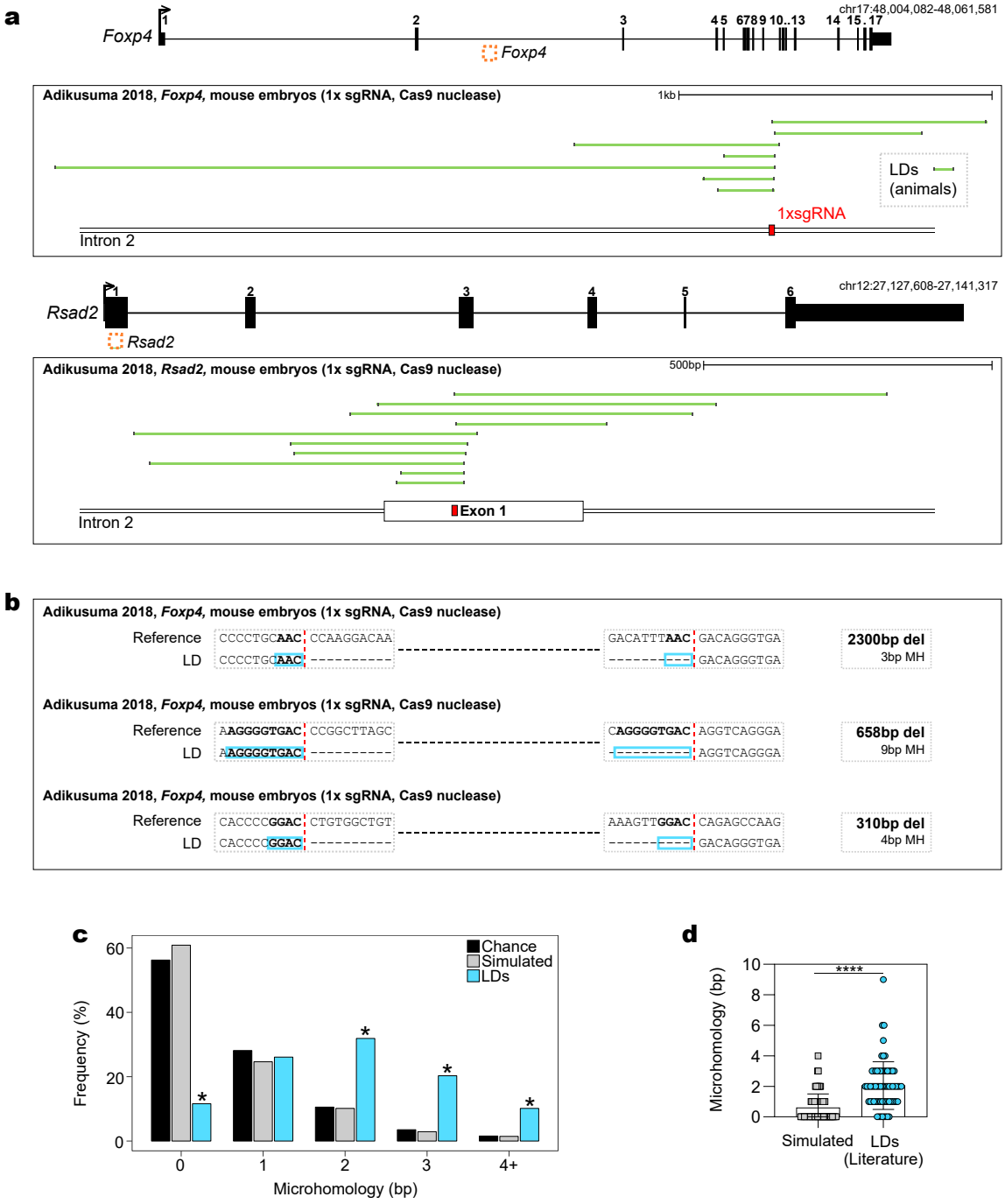


Figure 7.13 – Previously published larger deletion alleles analysed for the presence of microhomologies. *Legend continued on next page.*

Figure 7.13 – *Legend continued from previous page.* a) Representative previously published (Adikusuma et al., 2018) LD alleles (green lines) identified at two selected sites targeted with 1xsgRNA (red bars) and Cas9 nuclease. c) Selected LD alleles previously published (Adikusuma et al., 2018) and the amount of microhomology that could be identified. Examples of LDs with microhomologies and corresponding reference sequences shown (mm9). Sequences outlined with blue boxes and highlighted in bold represent microhomologies. Red dashed vertical lines represent the exact breakpoint junctions in the repaired alleles. Total deletion size and microhomology amount are indicated. c) Frequency distribution of microhomology amount found in 69 previously published LDs (Wang et al., 2013; Zhou et al., 2014; Ma et al., 2014; Parikh et al., 2015; Zhang et al., 2015; Mianne et al., 2017; Adikusuma et al., 2018) that were analysed for the presence of microhomology, the same number of simulated deletions across the genome with the same average length as LDs (Simulated), and the expected probability of finding homology at two sites for a k-mer of a given length (Chance) (*, chi2 test, $p < 7^{-9}$). d) Quantification of the amount of microhomology identified at 69 previously published LD alleles and simulated deletions (****, Mann-Whitney test, $p < 0.0001$).

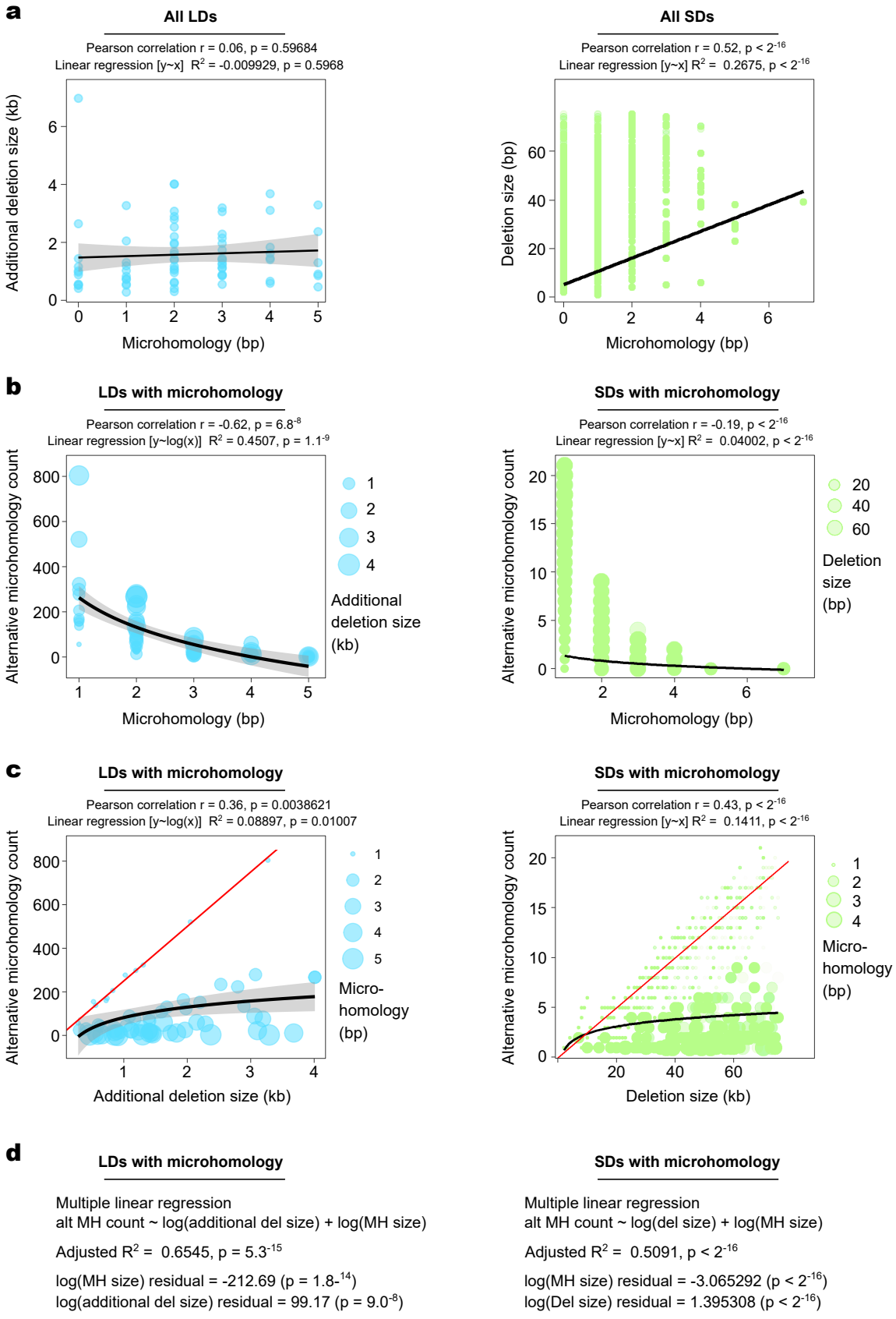
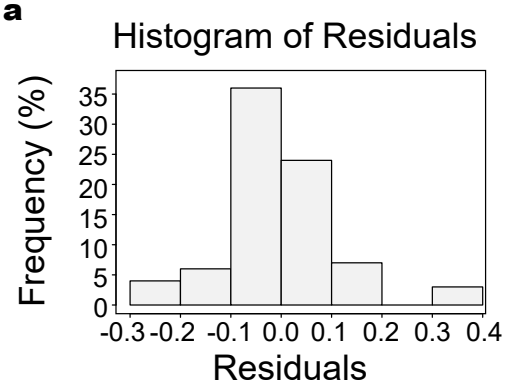


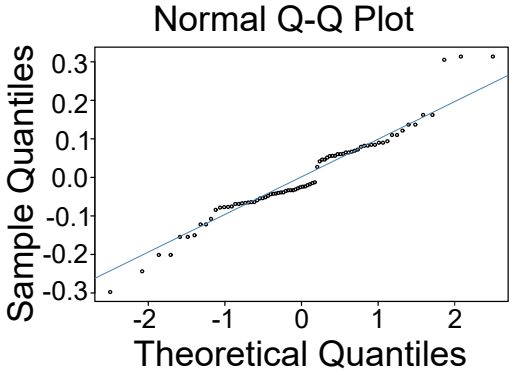
Figure 7.14 – Distribution of microhomology sequences is dependent on microhomology length and deletion size. *Legend continued on next page.*

Figure 7.14 – *Legend continued from previous page.* a) Scatter plot showing deletion size vs microhomology size at all LDs and all SDs at *Runx1* dsCTCF-3 and *Prickle2* exon 6. Linear regression was done using the given formula and shown on the plot with 95% confidence intervals. R^2 and p values are indicated above the plot with Pearson correlation r and p values. b) Scatter plot of alternative microhomology count vs microhomology size at all LDs and all SDs at *Runx1* dsCTCF-3 and *Prickle2* exon 6 with microhomology. Linear regression was done using the given formula and shown on the plot with 95% confidence intervals. R^2 and p values are indicated above the plot with Pearson correlation r and p values. c) Scatter plot of alternative microhomology count vs deletion size at all LDs and all SDs at *Runx1* dsCTCF-3 and *Prickle2* exon 6 with microhomology. The red line (y=x/4) indicates the chance of finding one nucleotide in a stretch of DNA with evenly distributed nucleotides of length x. Linear regression was done using the given formula and is shown on the plot with 95% confidence intervals. R^2 and p values are indicated above the plot with Pearson correlation r and p values. d) Multiple linear regression was done using the given formula with adjusted R^2 with p values, and residuals with p values indicated. The higher R^2 values in d compared with b and c indicates that alternative microhomology count is dependent on the combination of both deletion size and microhomology size.

a



b



c

Multiple linear regression summary

```
Call:
lm(formula = deletionFrequency ~ log(proximity) + cutEfficiency, data = All)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21981 -0.06501 -0.02923  0.06411  0.30111

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.624518   0.070606   8.845 1.57e-12 ***
log(proximity)  -0.103572   0.005979 -17.323  < 2e-16 ***
cutEfficiency    0.250865   0.083411   3.008  0.00382 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1067 on 61 degrees of freedom
Multiple R-squared:  0.833,  Adjusted R-squared:  0.8275
F-statistic: 152.1 on 2 and 61 DF,  p-value: < 2.2e-16
```

Figure 7.15 – Multiple linear regression model of the distribution of Cas9-induced larger deletion sizes. a) Histogram of residuals for the model (where residual = observed value - predicted value). b) Normal Q-Q plot for the model to check that residuals are normally distributed (which is an assumption of the linear regression model). c) Summary of multiple linear regression using the given formula with adjusted R^2, p values, and residuals with p values indicated.

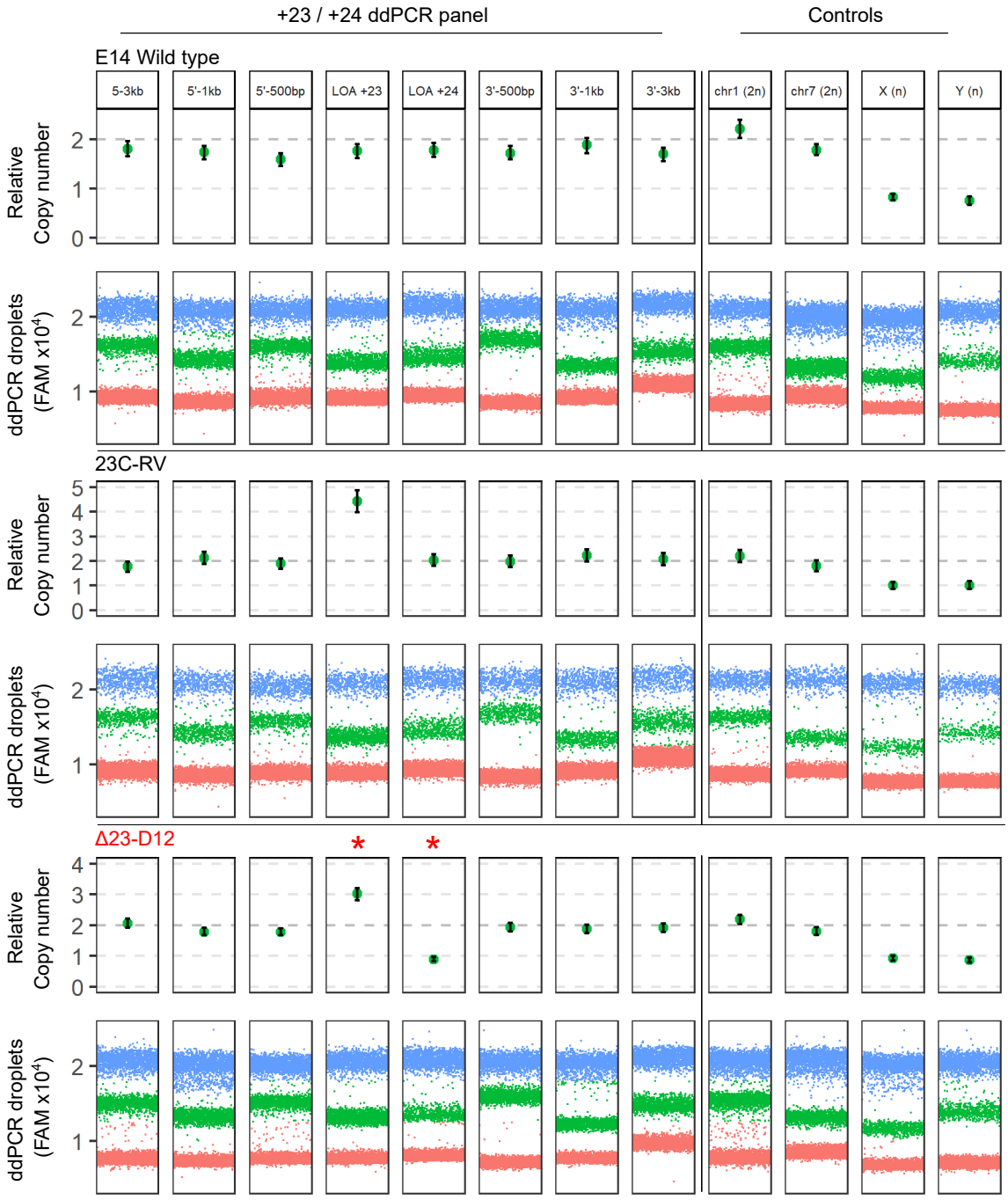


Figure 7.16 – Raw droplet digital PCR analysis at of +23 enhancer knock-out clones.
Legend continued on next page.

Figure 7.16 – *Legend continued from previous page.* ddPCR genotyping over the +23/+24 enhancer region from E14 wild type (top two rows), 23C-RV (middle two rows), and clone Δ23-D12. ddPCR droplets used to calculate relative copy numbers are indicated in the bottom row for each clone. Droplets were gated based on fluorescence of EvaGreen® that allowed distinction between a 100 bp test amplicon (present in the green droplets) and a 200 bp internal control amplicon (present in the blue droplets). Red droplets were negative for both templates. Relative copy numbers indicated (top row for each clone) were calculated relative to the control amplicon in each well, and to two different diploid chromosomes (chr1 and chr7) that were not targeted using $Cas9^{D10A}$ nickase. Amplicons targeting chrX and chrY were included as single copy controls. E14-Wt cells were used as a diploid control (relative copy number 2) across all expected amplicons and a haploid control (relative copy number 1) for the sex chromosomes. 23C-RV cells were included as a 4n control as they were previously targeted with a +23-Cherry enhancer reporter construct and are homozygous for this knock-in (so relative copy number = 4 in total). One additional copy of *Runx1* +23/+24 enhancers are indicated with red stars above the copy number plots.

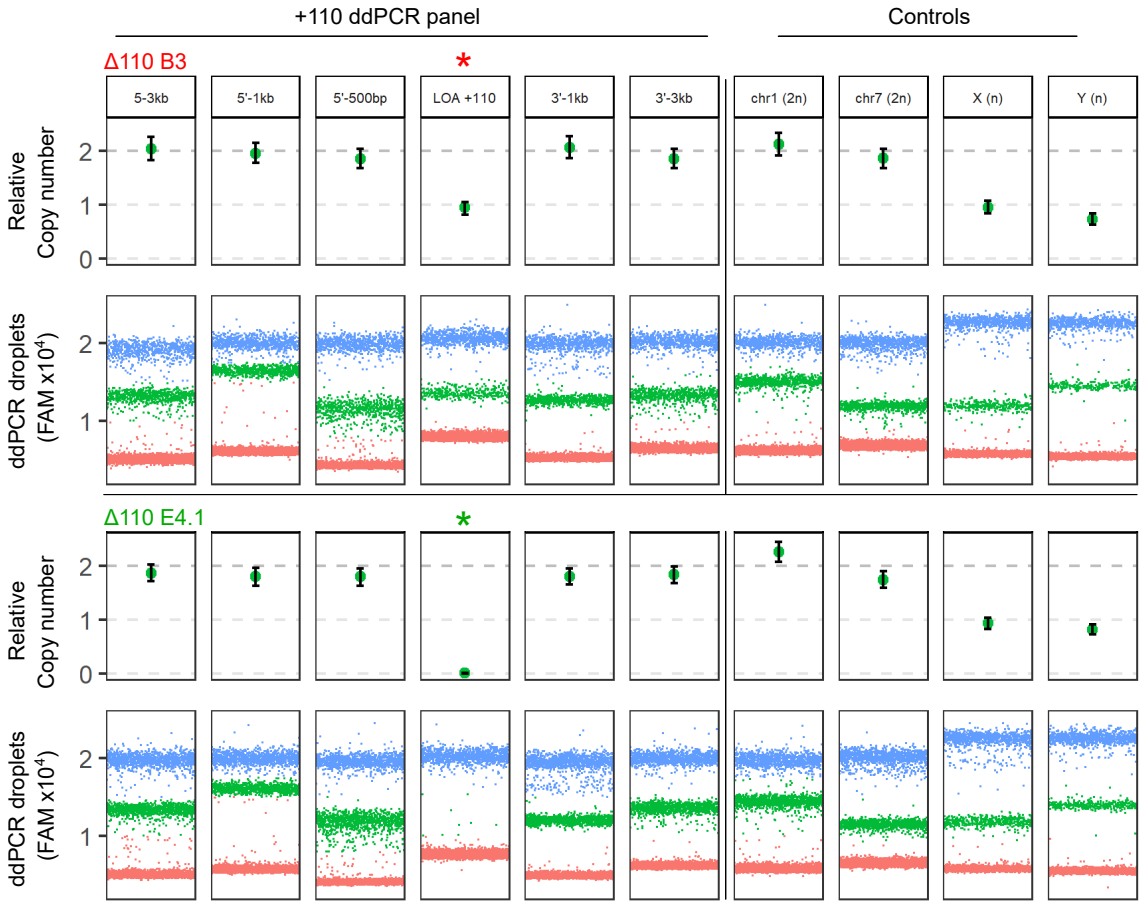


Figure 7.17 – Raw droplet digital PCR analysis at of +110 enhancer knock-out clones. *Legend continued on next page.*

Figure 7.17 – *Legend continued from previous page.* ddPCR genotyping over the +110 enhancer region from Δ110-B3 (top two rows), and Δ110-E4.1 (bottom two rows). ddPCR droplets used to calculate relative copy numbers are indicated in the bottom row for each clone. Droplets were gated based on fluorescence of EvaGreen® that allowed distinction between a 100 bp test amplicon (present in the green droplets) and a 200 bp internal control amplicon (present in the blue droplets). Red droplets were negative for both templates. Relative copy numbers indicated (top row for each clone) were calculated relative to the control amplicon in each well, and to two different diploid chromosomes (chr1 and chr7) that were not targeted using $Cas9^{D10A}$ nickase. Amplicons targeting chrX and chrY were included as single copy controls. One copy of *Runx1* +110 enhancer that was retained in clone Δ110-B3 is indicated with a red star above the copy number plot. Loss of *Runx1* +110 enhancer in clone Δ110-E4.1 is indicated with a green star above the copy number plot.

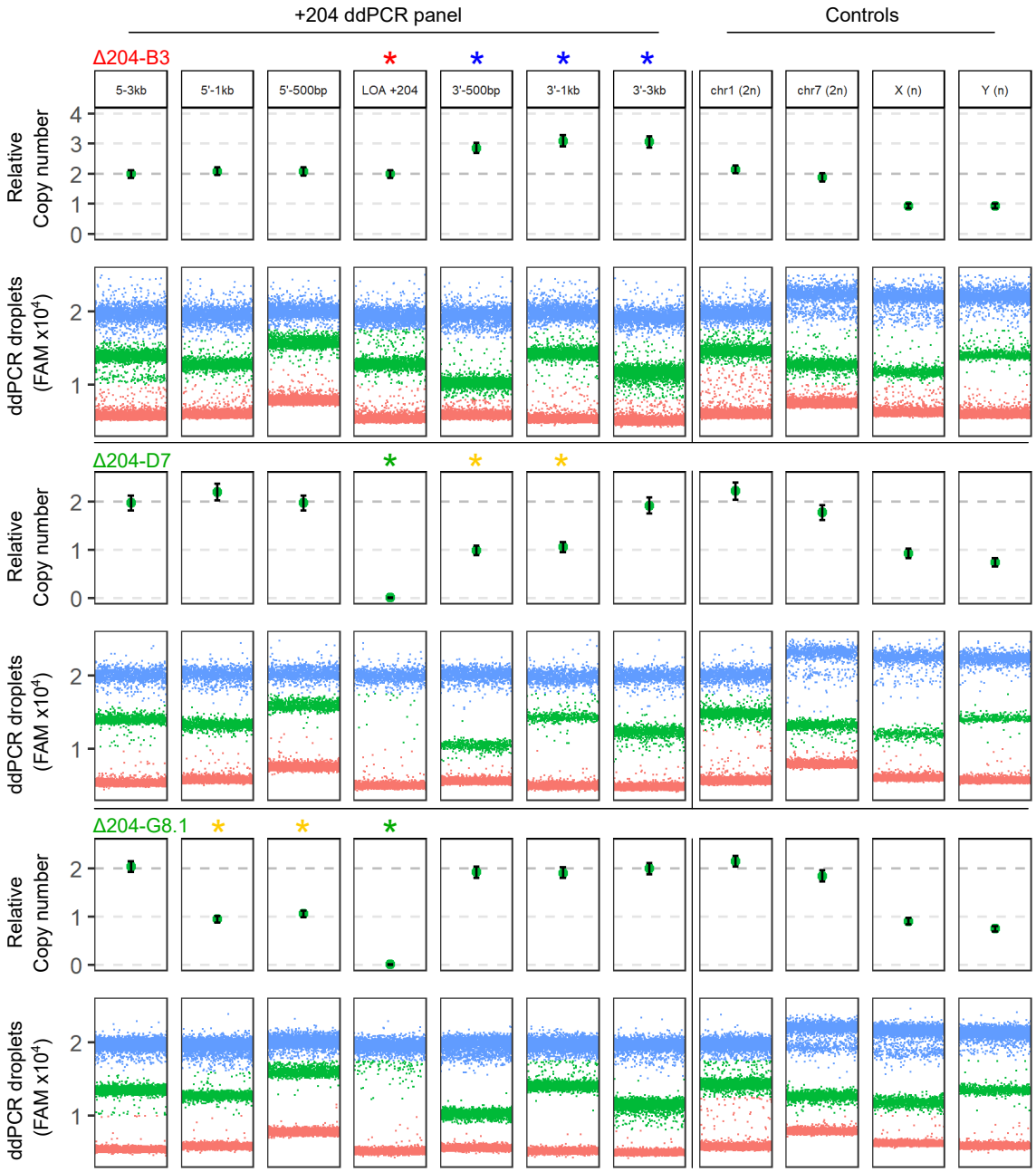


Figure 7.18 – Raw droplet digital PCR analysis at of +204 enhancer knock-out clones.
Legend continued on next page.

Figure 7.18 – *Legend continued from previous page.* ddPCR genotyping over the +204 enhancer region from Δ204-B3 (top two rows), Δ204-D7 (middle two rows), and Δ204-G8.1 (bottom two rows). ddPCR droplets used to calculate relative copy numbers are indicated in the bottom row for each clone. Droplets were gated based on fluorescence of EvaGreen® that allowed distinction between a 100 bp test amplicon (present in the green droplets) and a 200 bp internal control amplicon (present in the blue droplets). Red droplets were negative for both templates. Relative copy numbers indicated (top row for each clone) were calculated relative to the control amplicon in each well, and to two different diploid chromosomes (chr1 and chr7) that were not targeted using $Cas9^{D10A}$ nickase. Amplicons targeting chrX and chrY were included as single copy controls. Three copies of the region 3' to *Runx1* +204 enhancer that were quantified in clone Δ204-B3 are indicated with blue stars above the copy number plots. Two copies of the +204 enhancer that were seen in clone Δ204-B3 is indicated with a red star above the copy number plot. Loss of *Runx1* +204 enhancer in clones Δ204-D7 and Δ204-G8.1 are indicated with a green star above the copy number plots. LDs that likely destroyed primer binding sites in clones Δ204-D7 and Δ204-G8.1 are indicated with orange stars above the copy number plots.

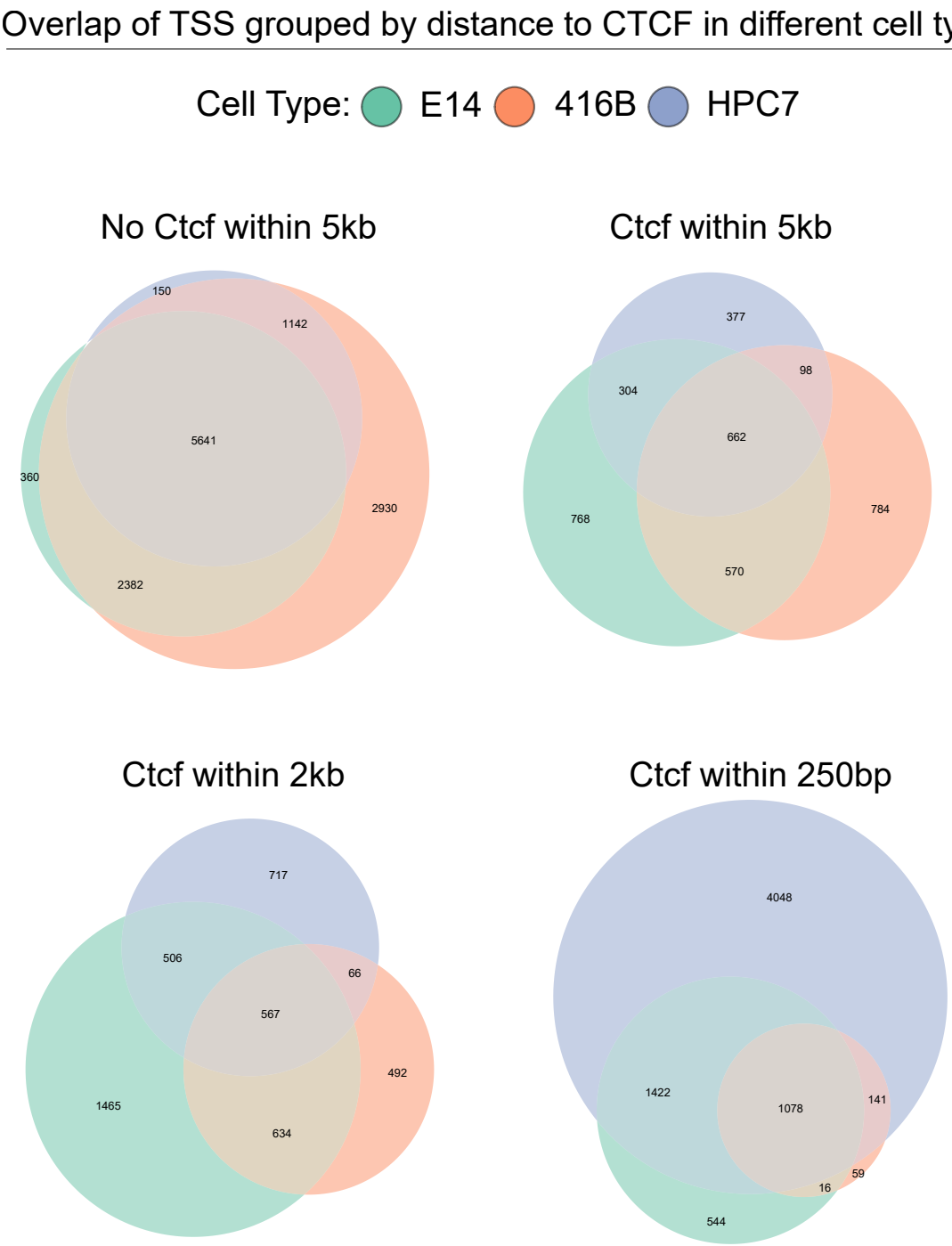


Figure 7.19 – Overlap in different cell types of TSS grouped by distance to CTCF peaks. 17263 non-overlapping TSS from the refGene database annotated by distance to CTCF peak. TSSs were grouped by binding of CTCF within a 5kb/2kb/250bp window centered on the annotated TSS, or no binding within a 5kb window.

Table 7.1 – Hubs of sequencing data generated for this thesis that can be loaded into the UCSC genome browser.

Hub info
416B Capture-C http://sara.molbiol.ox.ac.uk/public/dowens/CapC_CM5_autoHubR/hub_B416/B416_hub.txt
E14 Capture-C http://sara.molbiol.ox.ac.uk/public/dowens/CapC_CM5_autoHubR/hub_E14/E14_hub.txt
416B Virtual Capture-C http://sara.molbiol.ox.ac.uk/public/dowens/CapC_CM5_virtualCapC/hub_B416/B416_hub.txt
E14 Virtual Capture-C http://sara.molbiol.ox.ac.uk/public/dowens/CapC_CM5_virtualCapC/hub_E14/E14_hub.txt
416B Rad21 ChIP-seq http://userweb.molbiol.ox.ac.uk/public/dowens/Rad21_ChIP/hub/Rad21_ChIP/HubFolder/hub.txt
E14 Rad21 ChIP-seq http://userweb.molbiol.ox.ac.uk/public/dowens/Rad21_ChIP_E14/hub/Rad21_ChIP_E14/HubFolder/hub.txt
416B CTCF ChIP-seq http://userweb.molbiol.ox.ac.uk/public/dowens/416B_Ctcf_merge/hub/416B_Ctcf_merge/HubFolder/hub.txt
HE & HP ATAC-seq http://sara.molbiol.ox.ac.uk/public/dowens/ATAC_mESC_final/hub.txt
E14 & 416B & HPC7 *Runx1* P1 & P2 Bisulfite-seq http://sara.molbiol.ox.ac.uk/public/dowens/Methylation/hub.txt
E14 & 416B & HPC7 Poly A plus and minus RNA-seq http://sara.molbiol.ox.ac.uk/public/dowens/RNA-seq_final/hub.txt

Table 7.2 – UCSC hubs of public sequencing data analysed in this thesis

Hub info
E14 DNaseI-seq (Vierstra et al., 2014)
http://userweb.molbiol.ox.ac.uk/public/dowens/mESC_E14_hub/hub/HubFolder/hub.txt
E14 H3K27ac ChIP-seq (Wamstad et al., 2012)
http://userweb.molbiol.ox.ac.uk/public/dowens/E14_histones/hub/E14_H3K27ac/HubFolder/hub.txt
E14 H3K4me1 ChIP-seq (Wamstad et al., 2012)
http://userweb.molbiol.ox.ac.uk/public/dowens/E14_histones/hub/E14_H3K4me1/HubFolder/hub.txt
E14 CTCF ChIP-seq (Handoko et al., 2011)
http://userweb.molbiol.ox.ac.uk/public/dowens/E14_Handoko/hub/E14_Ctcf_Handoko/HubFolder/hub.txt
E14 Pol II ChIP-seq (Rahl et al., 2010)
http://userweb.molbiol.ox.ac.uk/public/dowens/mES_V6.5_Pol2_Rahl/hub/mES_V6.5_Pol2_Rahl/HubFolder/hub.txt
416B DNaseI-seq (Vierstra et al., 2014)
http://userweb.molbiol.ox.ac.uk/public/dowens/Stam_2012/hub/HubFolder/hub.txt
416B DNaseI-seq with footprinting (Vierstra et al., 2014)
http://userweb.molbiol.ox.ac.uk/public/dowens/Stam_2012_ftpt/hub/416B_DNase_ftpt/HubFolder/hub.txt
416B transcription factors and H3K27ac ChIP-seq (Schütte et al., 2016)
http://userweb.molbiol.ox.ac.uk/public/dowens/416B_Schutte_2016/hub/HubFolder/hub.txt
416B TFs and H3K27ac ChIP-seq (Schütte et al., 2016)
http://userweb.molbiol.ox.ac.uk/public/dowens/416B_Schutte_2016/hub/HubFolder/hub.txt
HPC7 DNaseI-seq (Wilson et al., 2010a)
http://userweb.molbiol.ox.ac.uk/public/dowens/HPC7_DNase/hub/HubFolder/hub.txt
HPC7 DNaseI-seq with footprinting (Wilson et al., 2010a)
http://userweb.molbiol.ox.ac.uk/public/dowens/HPC7_DNase_ftpt/hub/HPC7_DNase_ftpt/HubFolder/hub.txt
HPC7 H3K27ac and CTCF ChIP-seq (Calero-Nieto et al., 2014)
http://userweb.molbiol.ox.ac.uk/public/dowens/E14_histones/hub/E14_H3K27ac/HubFolder/hub.txt
HPC7 Rad21 ChIP-seq (Wilson et al., 2016)
http://userweb.molbiol.ox.ac.uk/public/dowens/HPC7_Wilson_2016_v2/hub/HubFolder/hub.txt

References

Abedin, M. J., Nguyen, A., Jiang, N., Perry, C. E., Shelton, J. M., Watson, D. K., and Ferdous, A. (2014). Fli1 acts downstream of etv2 to govern cell survival and vascular homeostasis via positive autoregulation. *Circulation research*, 114(11):1690–1699.

Adikusuma, F., Piltz, S., Corbett, M. A., Turvey, M., McColl, S. R., Helbig, K. J., Beard, M. R., Hughes, J., Pomerantz, R. T., and Thomas, P. Q. (2018). Large deletions induced by Cas9 cleavage. *Nature*, 560(7717):E8–E9.

Ahituv, N., Prabhakar, S., Poulin, F., Rubin, E. M., and Couronne, O. (2005). Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. *Human molecular genetics*, 14(20):3057–3063.

Akhtar, W., de Jong, J., Pindyurin, A. V., Pagie, L., Meuleman, W., de Ridder, J., Berns, A., Wessels, L. F., van Lohuizen, M., and van Steensel, B. (2013). Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell*, 154(4):914–927.

Alipour, E. and Marko, J. F. (2012). Self-organization of domain structures by dna-loop-extruding enzymes. *Nucleic Acids Res*, 40(22):11202–12.

Allen, B. L. and Taatjes, D. J. (2015). The mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol*, 16(3):155–66.

Allen, F., Crepaldi, L., Alsinet, C., Strong, A. J., Kleshchevnikov, V., De Angeli, P., Palenikova, P., Khodak, A., Kiselev, V., Kosicki, M., Bassett, A. R., Harding, H., Galanty, Y., Munoz-Martinez, F., Metzakopian, E., Jackson, S. P., and Parts, L. (2018). Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.*

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (2009). Chromosomal dynamics at the shh locus: limb bud-specific differential regulation of competence and active transcription. *Developmental cell*, 16(1):47–57.

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., Trono, D., Spitz, F., and Duboule, D. (2013). A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science*, 340(6137):1234167.

Anguita, E., Sharpe, J. A., Sloane-Stanley, J. A., Tufarelli, C., Higgs, D. R., and Wood, W. G. (2002). Deletion of the mouse alpha-globin regulatory element (HS -26) has an unexpectedly mild phenotype. *Blood*, 100(10):3450–3456.

Appel, E., Weissmann, S., Salzberg, Y., Orlovsky, K., Negreanu, V., Tsoory, M., Raanan, C., Feldmesser, E., Bernstein, Y., Wolstein, O., et al. (2016). An ensemble of regulatory elements controls runx3 spatiotemporal expression in subsets of dorsal root ganglia proprioceptive neurons. *Genes & development*, 30(23):2607–2622.

Arab, K., Karaulanov, E., Musheev, M., Trnka, P., Schafer, A., Grummt, I., and Niehrs, C. (2019). GADD45A binds R-loops and recruits TET1 to CpG island promoters. *Nat. Genet.*, 51(2):217–223.

Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339(6123):1074–1077.

Arzate-Mejia, R. G., Recillas-Targa, F., and Corces, V. G. (2018). Developing in 3D: the role of CTCF in cell differentiation. *Development*, 145(6).

Ata, H., Ekstrom, T. L., Martinez-Galvez, G., Mann, C. M., Dvornikov, A. V., Schaefbauer, K. J., Ma, A. C., Dobbs, D., Clark, K. J., and Ekker, S. C. (2018). Robust activation of microhomology-mediated end joining for precision gene editing applications. *PLoS Genet.*, 14(9):e1007652.

Azcoitia, V., Aracil, M., Martínez-A, C., and Torres, M. (2005). The homeodomain protein meis1 is essential for definitive hematopoiesis and vascular patterning in the mouse embryo. *Developmental biology*, 280(2):307–320.

Bae, S., Kweon, J., Kim, H. S., and Kim, J. S. (2014). Microhomology-based choice of Cas9 nuclease target sites. *Nat. Methods*, 11(7):705–706.

Bahr, C., von Paleske, L., Uslu, V. V., Remeseiro, S., Takayama, N., Ng, S. W., Murison, A., Langenfeld, K., Petretich, M., Scognamiglio, R., Zeisberger, P., Benk, A. S., Amit, I., Zandstra, P. W., Lupien, M., Dick, J. E., Trumpp, A., and Spitz, F. (2018). A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature*, 553(7689):515–520.

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36.

Bancaud, A., L. C. H. S. and Ellenberg, J. (2012). A fractal model for nuclear organization: current evidence and biological implications. *Nucleic Acids Res*, 40(18):8783–92.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote sv40 dna sequences. *Cell*, 27(2):299–308.

Banerji, J., O. L. and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, 33(3):729–40.

Barrington, C., Georgopoulou, D., Pezic, D., Varsally, W., Herrero, J., and Hadjur, S. (2019). Enhancer accessibility and CTCF occupancy underlie asymmetric TAD architecture and cell type specific genome topology. *Nat Commun*, 10(1):2908.

Baubec, T., Colombo, D. F., Wirbelauer, C., Schmidt, J., Burger, L., Krebs, A. R., Akalin, A., and Schubeler, D. (2015). Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature*, 520(7546):243–247.

Beagan, J. A., Duong, M. T., Titus, K. R., Zhou, L., Cao, Z., Ma, J., Lachanski, C. V., Gillis, D. R., and Phillips-Cremins, J. E. (2017). YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res.*, 27(7):1139–1152.

Bee, T., Ashley, E. L., Bickley, S. R., Jarratt, A., Li, P. S., Sloane-Stanley, J., Gottgens, B., and de Bruijn, M. F. (2009a). The mouse Runx1 +23 hematopoietic stem cell enhancer confers hematopoietic specificity to both Runx1 promoters. *Blood*, 113(21):5121–5124.

Bee, T., Liddiard, K., Swiers, G., Bickley, S. R., Vink, C. S., Jarratt, A., Hughes, J. R., Medvinsky, A., and de Bruijn, M. F. (2009b). Alternative runx1 promoter usage in mouse developmental hematopoiesis. *Blood Cells, Molecules, and Diseases*, 43(1):35–42.

Bee, T., Swiers, G., Muroi, S., Pozner, A., Nottingham, W., Santos, A. C., Li, P.-S., Taniuchi, I., and de Bruijn, M. F. (2010). Nonredundant roles for runx1 alternative promoters reflect their activity at discrete stages of developmental hematopoiesis. *Blood*, 115(15):3042–3050.

Bell, A. C. and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, 405(6785):482–485.

Bell, A. C., W. A. G. and Felsenfeld, G. (1999). The protein ctcf is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98(3):387–96.

Ben-Ami, O., Pencovich, N., Lotem, J., Levanon, D., and Groner, Y. (2009). A regulatory interplay between miR-27a and Runx1 during megakaryopoiesis. *Proc. Natl. Acad. Sci. U.S.A.*, 106(1):238–243.

Benabdallah, N. S., Williamson, I., Illingworth, R. S., Kane, L., Boyle, S., Sengupta, D., Grimes, G. R., Therizols, P., and Bickmore, W. A. (2019). Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Mol. Cell*.

Bender, M. A., Ragoczy, T., Lee, J., Byron, R., Telling, A., Dean, A., and Groudine, M. (2012). The hypersensitive sites of the murine Î²-globin locus control region act independently to affect nuclear localization and transcriptional elongation. *Blood*, 119(16):3820–3827.

Benner, C., Isoda, T., and Murre, C. (2015). New roles for DNA cytosine modification, eRNA, anchors, and superanchors in developing B cell progenitors. *Proc. Natl. Acad. Sci. U.S.A.*, 112(41):12776–12781.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326.

Bertrand, J. Y., Chi, N. C., Santoso, B., Teng, S., Stainier, D. Y., and Traver, D. (2010). Haematopoietic stem cells derive directly from aortic endothelium during development. *Nature*, 464(7285):108–111.

Bickmore, W. A. and van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. *Cell*, 152(6):1270–1284.

Birling, M. C., Schaeffer, L., Andre, P., Lindner, L., Marechal, D., Ayadi, A., Sorg, T., Pavlovic, G., and Herault, Y. (2017). Efficient and rapid generation of large genomic variants in rats and mice using CRISMERE. *Sci Rep*, 7:43331.

Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganière, J., Lefèbvre, C., Deblois, G., Giguère, V., Ferretti, V., Bergeron, D., et al. (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome research*, 16(5):656–668.

Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2010). Chip-seq identification of weakly conserved heart enhancers. *Nature genetics*, 42(9):806.

Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A. S., Yu, T., Marie-Nelly, H., McSwiggen, D. T., Kokic, G., Dailey, G. M., Cramer, P., Darzacq, X., and Zweckstetter, M. (2018). RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nat. Struct. Mol. Biol.*, 25(9):833–840.

Boisset, J.-C., van Cappellen, W., Andrieu-Soler, C., Galjart, N., Dzierzak, E., and Robin, C. (2010). In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium. *Nature*, 464(7285):116–120.

Bonev, B. and Cavalli, G. (2016). Organization and function of the 3D genome. *Nat. Rev. Genet.*, 17(11):661–678.

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J. P., Tanay, A., and Cavalli, G. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, 171(3):557–572.

Boriushkin, E., Zhang, H., Becker, M., Peachey, J., Shatat, M. A., Adams, R. H., and Hamik, A. (2019). Kruppel-like factor 4 regulates developmental angiogenesis through disruption of the RBP-J-NICD-MAML complex in intron 3 of Dll4. *Angiogenesis*, 22(2):295–309.

Boroviak, K., Fu, B., Yang, F., Doe, B., and Bradley, A. (2017). Revealing hidden complexities of genomic rearrangements generated with Cas9. *Sci Rep*, 7(1):12867.

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J. L., Ruan, Y., Wei, C. L., Ng, H. H., and Liu, E. T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.*, 18(11):1752–1762.

Brenowitz, M., Senear, D. F., Shea, M. A., and Ackers, G. K. (1986). Quantitative DNase footprint titration: a method for studying protein-DNA interactions. *Meth. Enzymol.*, 130:132–181.

Brinkman, E. K., Chen, T., de Haas, M., Holland, H. A., Akhtar, W., and van Steensel, B. (2018). Kinetics and Fidelity of the Repair of Cas9-Induced Double-Strand DNA Breaks. *Mol. Cell*, 70(5):801–813.

Brown, J. M., Leach, J., Reittie, J. E., Atzberger, A., Lee-Prudhoe, J., Wood, W. G., Higgs, D. R., Iborra, F. J., and Buckle, V. J. (2006). Coregulated human globin genes are frequently in spatial proximity when active. *J. Cell Biol.*, 172(2):177–187.

Brown, J. M., Roberts, N. A., Graham, B., Waithe, D., Lagerholm, C., Telenius, J. M., De Ornellas, S., Oudelaar, A. M., Scott, C., Szczerbal, I., Babbs, C., Kassouf, M. T., Hughes, J. R., Higgs, D. R., and Buckle, V. J. (2018). A tissue-specific self-interacting chromatin domain forms independently of enhancer-promoter interactions. *Nat Commun*, 9(1):3849.

Buecker, C. and Wysocka, J. (2012). Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet.*, 28(6):276–284.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213.

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). Atac-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, pages 21–29.

Burns, C. E., Traver, D., Mayhall, E., Shepard, J. L., and Zon, L. I. (2005). Hematopoietic stem cell fate is established by the notch–runx pathway. *Genes & development*, 19(19):2331–2342.

Bushman, F. D. (2003). Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. *Cell*, 115(2):135–138.

Cai, Z., de Bruijn, M., Ma, X., Dortland, B., Luteijn, T., Downing, J., and Dzierzak, E. (2000). Haploinsufficiency of aml1/cbfa2 affects the embryonic generation of mouse hematopoietic stem cells. *Immunity*, 13(4):423–431.

Calero-Nieto, F. J., Ng, F. S., Wilson, N. K., Hannah, R., Moignard, V., Leal-Cervantes, A. I., Jimenez-Madrid, I., Diamanti, E., Wernisch, L., and Göttgens, B. (2014). Key regulators control distinct transcriptional programmes in blood progenitor and mast cells. *The EMBO journal*, 33(11):1212–1226.

Cannavò, E., Khoueiry, P., Garfield, D. A., Geeleher, P., Zichner, T., Gustafson, E. H., Ciglar, L., Korbel, J. O., and Furlong, E. E. (2016). Shadow enhancers are pervasive features of developmental regulatory networks. *Current Biology*, 26(1):38–51.

Canver, M. C., Bauer, D. E., Dass, A., Yien, Y. Y., Chung, J., Masuda, T., Maeda, T., Paw, B. H., and Orkin, S. H. (2014). Characterization of genomic deletion efficiency mediated by clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells. *J. Biol. Chem.*, 289(31):21312–21324.

Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., Chen, D. D., Schupp, P. G., Vinjamur, D. S., Garcia, S. P., et al. (2015). Bcl11a enhancer dissection by cas9-mediated in situ saturating mutagenesis. *Nature*.

Canzio, D., Nwakeze, C. L., Horta, A., Rajkumar, S. M., Coffey, E. L., Duffy, E. E., Duffie, R., Monahan, K., O'Keeffe, S., Simon, M. D., Lomvardas, S., and Maniatis, T. (2019). Antisense lncRNA Transcription Mediates DNA Demethylation to Drive Stochastic Protocadherin Î± Promoter Choice. *Cell*, 177(3):639–653.

Cao, Z., Sun, X., Icli, B., Wara, A. K., and Feinberg, M. W. (2010). Role of Kruppel-like factors in leukocyte development, function, and disease. *Blood*, 116(22):4404–4414.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstrom, P. G., Frith, M. C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics*, 38(6):626.

Catarino, R. R. and Stark, A. (2018). Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev.*, 32(3-4):202–223.

Chakrabarti, A. M., Henser-Brownhill, T., Monserrat, J., Poetsch, A. R., Luscombe, N. M., and Scaffidi, P. (2019). Target-Specific Precision of CRISPR-Mediated Genome Editing. *Mol. Cell*, 73(4):699–713.

Challen, G. A. and Goodell, M. A. (2010). Runx1 isoforms show differential expression patterns during hematopoietic development but have similar functional effects in adult hematopoietic stem cells. *Exp. Hematol.*, 38(5):403–416.

Chen, F. X., Xie, P., Collings, C. K., Cao, K., Aoi, Y., Marshall, S. A., Rendleman, E. J., Ugarenko, M., Ozark, P. A., Zhang, A., et al. (2017). Paf1 regulation of promoter-proximal pause release via enhancer activation. *Science*, 357(6357):1294–1298.

Chen, M. J., Yokomizo, T., Zeigler, B. M., Dzierzak, E., and Speck, N. A. (2009). Runx1 is required for the endothelial to haematopoietic cell transition but not thereafter. *Nature*, 457(7231):887–891.

Cho, W. K., Spille, J. H., Hecht, M., Lee, C., Li, C., Grube, V., and Cisse, I. I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, 361(6400):412–415.

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G. M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X., and Tjian, R. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science*, 361(6400).

Chung, J. H., Bell, A. C., and Felsenfeld, G. (1997). Characterization of the chicken beta-globin insulator. *Proc. Natl. Acad. Sci. U.S.A.*, 94(2):575–580.

Chung, J. H., Whiteley, M., and Felsenfeld, G. (1993). A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, 74(3):505–514.

Codner, G. F., Lindner, L., Caulder, A., Wattenhofer-Donze, M., Radage, A., Mertz, A., Eisenmann, B., Mianne, J., Evans, E. P., Beechey, C. V., Fray, M. D., Birling, M. C., Herault, Y., Pavlovic, G., and Teboul, L. (2016). Aneuploidy screening of embryonic stem cell clones by metaphase karyotyping and droplet digital polymerase chain reaction. *BMC Cell Biol.*, 17(1):30.

Codner, G. F., Mianne, J., Caulder, A., Loeffler, J., Fell, R., King, R., Allan, A. J., Mackenzie, M., Pike, F. J., McCabe, C. V., Christou, S., Joynson, S., Hutchison, M., Stewart, M. E., Kumar, S., Simon, M. M., Agius, L., Anstee, Q. M., Volynski, K. E., Kullmann, D. M., Wells, S., and Teboul, L. (2018). Application of long single-stranded DNA donors in genome editing: generation and validation of mouse mutants. *BMC Biol.*, 16(1):70.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., and Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121):819–823.

Cradick, T. J., Fine, E. J., Antico, C. J., and Bao, G. (2013). CRISPR/Cas9 systems targeting Î²-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.*, 41(20):9584–9592.

Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. A., Ginsburg, D., et al. (2006). Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpss). *Genome research*, 16(1):123–131.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., et al. (2010). Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936.

Curradi, M., Izzo, A., Badaracco, G., and Landsberger, N. (2002). Molecular mechanisms of gene silencing mediated by DNA methylation. *Mol. Cell. Biol.*, 22(9):3157–3173.

Danino, Y. M., Even, D., Ideses, D., and Juven-Gershon, T. (2015). The core promoter: At the heart of gene expression. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1849(8):1116–1131.

Darwin, C. (1872). *The Origin of Species: By Means of Natural Selection Or the Preservation of Favored Races in the Struggle for Life*, volume 1. Modern library.

Davalos-Salas, M., Furlan-Magaril, M., Gonzalez-Buendia, E., Valdes-Quezada, C., Ayala-Ortega, E., and Recillas-Targa, F. (2011). Gain of DNA methylation is enhanced in the absence of CTCF at the human retinoblastoma gene promoter. *BMC Cancer*, 11:232.

Davidson, I. F., Bauer, B., Goetz, D., Tang, W., Wutz, G., and Peters, J. M. (2019). DNA loop extrusion by human cohesin. *Science*, 366(6471):1338–1345.

Davies, J. O., Telenius, J. M., McGowan, S. J., Roberts, N. A., Taylor, S., Higgs, D. R., and Hughes, J. R. (2015). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature methods*.

de Bruijn, M. and Dzierzak, E. (2017). Runx transcription factors in the development and function of the definitive hematopoietic system. *Blood*, 129(15):2061–2069.

de Bruijn, M. F., Speck, N. A., Peeters, M. C., and Dzierzak, E. (2000). Definitive hematopoietic stem cells first develop within the major arterial regions of the mouse embryo. *The EMBO journal*, 19(11):2465–2474.

de Laat, W. and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, 502(7472):499–506.

de Wit, E., Vos, E. S., Holwerda, S. J., Valdes-Quezada, C., Verstegen, M. J., Teunissen, H., Splinter, E., Wijchers, P. J., Krijger, P. H., and de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. *Mol. Cell*, 60(4):676–684.

Deaton, A. M. and Bird, A. (2011). Cpg islands and the regulation of transcription. *Genes & development*, 25(10):1010–1022.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *science*, 295(5558):1306–1311.

Delgado-Olguin, P., Brand-Arzamendi, K., Scott, I. C., Jungblut, B., Stainier, D. Y., Bruneau, B. G., and Recillas-Targa, F. (2011). CTCF promotes muscle differentiation by modulating the activity of myogenic regulatory factors. *J. Biol. Chem.*, 286(14):12483–12494.

Deltcheva, E. and Nimmo, R. (2017). RUNX transcription factors at the interface of stem cells and cancer. *Biochem. J.*, 474(11):1755–1768.

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P. D., Dean, A., and Blobel, G. A. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, 149(6):1233–1244.

Dexter, T. M., Allen, T. D., Scott, D., and Teich, N. (1979). Isolation and characterisation of a bipotential haematopoietic cell line.

Ditadi, A., Sturgeon, C. M., and Keller, G. (2017). A view of human haematopoietic development from the Petri dish. *Nat. Rev. Mol. Cell Biol.*, 18(1):56–67.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. A., and Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.

Donohoe, M. E., Silva, S. S., Pinter, S. F., Xu, N., and Lee, J. T. (2009). The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting. *Nature*, 460(7251):128–132.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., and Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, 16(10):1299–1309.

Draper, J. E., Sroczynska, P., Leong, H. S., Fadlullah, M. Z. H., Miller, C., Kouskoff, V., and Lacaud, G. (2017). Mouse RUNX1C regulates premegakaryocytic/erythroid output and maintains survival of megakaryocyte progenitors. *Blood*, 130(3):271–284.

Driscoll, M. C., Dobkin, C. S., and Alter, B. P. (1989). Gamma delta beta-thalassemia due to a de novo mutation deleting the 5' beta-globin gene activation-region hypersensitive sites. *Proc. Natl. Acad. Sci. U.S.A.*, 86(19):7470–7474.

Dryden, N. H., Broome, L. R., Dudbridge, F., Johnson, N., Orr, N., Schoenfelder, S., Nagano, T., Andrews, S., Wingett, S., Kozarewa, I., Assiotis, I., Fenwick, K., Maguire, S. L., Campbell, J., Natrajan, R., Lambros, M., Perrakis, E., Ashworth, A., Fraser, P., and Fletcher, O. (2014). Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.*, 24(11):1854–1868.

Ebert, A., H. L. and Busslinger, M. (2015). Spatial regulation of v-(d)j recombination at antigen receptor loci. *Adv Immunol*, 128:93–121.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113.

Eijsbouts, C. Q., Burren, O. S., Newcombe, P. J., and Wallace, C. (2019). Fine mapping chromatin contacts in capture Hi-C data. *BMC Genomics*, 20(1):77.

Eilken, H. M., Nishikawa, S.-I., and Schroeder, T. (2009). Continuous single-cell imaging of blood generation from haemogenic endothelium. *Nature*, 457(7231):896–900.

Erdel, F. and Rippe, K. (2018). Formation of Chromatin Subcompartments by Phase Separation. *Biophys. J.*, 114(10):2262–2270.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43.

Espín-Palazón, R., Stachura, D. L., Campbell, C. A., García-Moreno, D., Del Cid, N., Kim, A. D., Candel, S., Meseguer, J., Mulero, V., and Traver, D. (2014). Proinflammatory signaling regulates hematopoietic stem cell emergence. *Cell*, 159(5):1070–1085.

Faure, A. J., Schmidt, D., Watt, S., Schwalie, P. C., Wilson, M. D., Xu, H., Ramsay, R. G., Odom, D. T., and Flicek, P. (2012). Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome Res.*, 22(11):2163–2175.

Feric, M., Vaidya, N., Harmon, T. S., Mitrea, D. M., Zhu, L., Richardson, T. M., Kriwacki, R. W., Pappu, R. V., and Brangwynne, C. P. (2016). Coexisting Liquid Phases Underlie Nucleolar Subcompartments. *Cell*, 165(7):1686–1697.

Filippova, D., P. R. D. G. and Kingsford, C. (2014). Identification of alternative topological domains in chromatin. *Algorithms Mol Biol*, 9:14.

Finver, S. N., Nishikura, K., Finger, L. R., Haluska, F. G., Finan, J., Nowell, P. C., and Croce, C. M. (1988). Sequence analysis of the MYC oncogene involved in the t(8;14)(q24;q11) chromosome translocation in a human leukemia T-cell line indicates that putative regulatory regions are not altered. *Proc. Natl. Acad. Sci. U.S.A.*, 85(9):3052–3056.

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., Suva, M. L., and Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, 529(7584):110–114.

Fonseca, G. J., Tao, J., Westin, E. M., Duttke, S. H., Spann, N. J., Strid, T., Shen, Z., Stender, J. D., Sakai, M., Link, V. M., Benner, C., and Glass, C. K. (2019). Diverse motif ensembles specify non-redundant DNA binding activities of AP-1 family members in macrophages. *Nat Commun*, 10(1):414.

Forcato, M., Nicoletti, C., Pal, K., Livi, C. M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. *Nat. Methods*, 14(7):679–685.

Forrester, W. C., Takegawa, S., Papayannopoulou, T., Stamatoyannopoulos, G., and Groudine, M. (1987). Evidence for a locus activation region: the formation of developmentally stable hypersensitive sites in globin-expressing hybrids. *Nucleic Acids Res.*, 15(24):10159–10177.

Fouse, S. D., Shen, Y., Pellegrini, M., Cole, S., Meissner, A., Van Neste, L., Jaenisch, R., and Fan, G. (2008). Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. *Cell Stem Cell*, 2(2):160–169.

Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schopflin, R., Kraft, K., Kempfer, R., Jerkovi?, I., Chan, W. L., Spielmann, M., Timmermann, B., Wittler, L., Kurth, I., Cambiaso, P., Zuffardi, O., Houge, G., Lambie, L., Brancati, F., Pombo, A., Vingron, M., Spitz, F., and Mundlos, S. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624):265–269.

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., and Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, 466(7305):490–493.

Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B. L., Kraemer, D. C., Aitken, S., et al. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular systems biology*, 11(12):852.

Fuda, N. J., Ardehali, M. B., and Lis, J. T. (2009). Defining mechanisms that regulate rna polymerase ii transcription in vivo. *Nature*, 461(7261):186.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep*, 15(9):2038–2049.

Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S., and Engreitz, J. M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*, 354(6313):769–773.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G., Huang, P. Y., Welboren, W. J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., Lee, Y. K., Karuturi, R. K., Herve, T., Bourque, G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W. K., Liu, E. T., Wei, C. L., Cheung, E., and Ruan, Y. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64.

Galas, D. J. and Schmitz, A. (1978). Dnaase footprinting a simple method for the detection of protein-dna binding specificity. *Nucleic acids research*, 5(9):3157–3170.

Gao, X., Johnson, K. D., Chang, Y.-I., Boyer, M. E., Dewey, C. N., Zhang, J., and Bresnick, E. H. (2013). Gata2 cis-element is required for hematopoietic stem cell generation in the mammalian embryo. *The Journal of experimental medicine*, 210(13):2833–2842.

Garner, M. M. and Revzin, A. (1986). The use of gel electrophoresis to detect and study nucleic acid— protein interactions. *Trends in Biochemical Sciences*, 11(10):395 – 396.

Ghavi-Helm, Y., Klein, F. A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E. E. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, 512(7512):96–100.

Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., Gregory, L., Lonie, L., Chew, A., Wei, C.-L., et al. (2010). Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity*, 32(3):317–328.

Ghozi, M., Bernstein, Y., Negreanu, V., Levanon, D., and Groner, Y. (1996). Expression of the human acute myeloid leukemia gene aml1 is regulated by two promoter regions. *Proceedings of the National Academy of Sciences*, 93(5):1935–1940.

Gibcus, J. H. and Dekker, J. (2013). The hierarchy of the 3d genome. *Mol Cell*, 49(5):773–82.

Gibson, D. G., Young, L., Chuang, R. Y., Venter, J. C., Hutchison, C. A., and Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, 6(5):343–345.

Gillies, S. D., Morrison, S. L., Oi, V. T., and Tonegawa, S. (1983). A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell*, 33(3):717–728.

Gilmour, J., Assi, S. A., Jaegle, U., Kulu, D., van de Werken, H., Clarke, D., Westhead, D. R., Philipsen, S., and Bonifer, C. (2014). A crucial role for the ubiquitously expressed transcription factor Sp1 at early stages of hematopoietic specification. *Development*, 141(12):2391–2401.

Giorgio, E., Robyr, D., Spielmann, M., Ferrero, E., Di Gregorio, E., Imperiale, D., Vaula, G., Stamoulis, G., Santoni, F., Atzori, C., et al. (2015). A large genomic deletion leads to enhancer adoption by the lamin b1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (adld). *Human molecular genetics*, 24(11):3143–3154.

Glover, L., Jun, J., and Horn, D. (2011). Microhomology-mediated deletion and gene conversion in African trypanosomes. *Nucleic Acids Res.*, 39(4):1372–1380.

Goode, D. K., Obier, N., Vijayabaskar, M., Lie-A-Ling, M., Lilly, A. J., Hannah, R., Lichtinger, M., Batta, K., Florkowska, M., Patel, R., et al. (2016). Dynamic gene regulatory networks drive hematopoietic specification and differentiation. *Developmental cell*, 36(5):572–587.

Grosveld, F., van Assendelft, G. B., Greaves, D. R., and Kollias, G. (1987). Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell*, 51(6):975–985.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., de Klein, A., Wessels, L., de Laat, W., and van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951.

Gunnell, A., Webb, H. M., Wood, C. D., McClellan, M. J., Wichaidit, B., Kempkes, B., Jenner, R. G., Osborne, C., Farrell, P. J., and West, M. J. (2016). Runx super-enhancer control through the notch pathway by epstein-barr virus transcription factors regulates b cell growth. *Nucleic acids research*, 44(10):4636–4650.

Guo, Y., Monahan, K., Wu, H., Gertz, J., Varley, K. E., Li, W., Myers, R. M., Maniatis, T., and Wu, Q. (2012). CTCF/cohesin-mediated DNA looping is required for protocadherin Î± promoter choice. *Proc. Natl. Acad. Sci. U.S.A.*, 109(51):21081–21086.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D. U., Jung, I., Wu, H., Zhai, Y., Tang, Y., Lu, Y., Wu, Y., Jia, Z., Li, W., Zhang, M. Q., Ren, B., Krainer, A. R., Maniatis, T., and Wu, Q. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*, 162(4):900–910.

Hadjur, S., W. L. M. R. N. K. C. B. S. S. T. F. P. F. A. G. and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated ifng locus. *Nature*, 460(7253):410–3.

Hah, N., Danko, C. G., Core, L., Waterfall, J. J., Siepel, A., Lis, J. T., and Kraus, W. L. (2011). A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*, 145(4):622–634.

Hainer, S. J., Bo?kovi?, A., McCannell, K. N., Rando, O. J., and Fazzio, T. G. (2019). Profiling of Pluripotency Factors in Single Cells and Early Embryos. *Cell*, 177(5):1319–1329.

Hale, A. T., Tian, H., Anih, E., Recio, F. O., Shatat, M. A., Johnson, T., Liao, X., Ramirez-Bergeron, D. L., Proweller, A., Ishikawa, M., and Hamik, A. (2014). Endothelial Kruppel-like factor 4 regulates angiogenesis and the Notch signaling pathway. *J. Biol. Chem.*, 289(17):12016–12028.

Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., and Taipale, J. (2006). Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, 124(1):47–59.

Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., Lee, C. W., Ye, C., Ping, J. L., Mulawadi, F., Wong, E., Sheng, J., Zhang, Y., Poh, T., Chan, C. S., Kunarso, G., Shahab, A., Bourque, G., Cacheux-Rataboul, V., Sung, W. K., Ruan, Y., and Wei, C. L. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, 43(7):630–638.

Handyside, A. H., O'Neill, G. T., Jones, M., and Hooper, M. L. (1989). Use of BRL-conditioned medium in combination with feeder layers to isolate a diploid embryonal stem cell line. *Rouxs Arch. Dev. Biol.*, 198(1):48–56.

Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R., and Darzacq, X. (2017). CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife*, 6.

Hanssen, L. L. P., Kassouf, M. T., Oudelaar, A. M., Biggs, D., Preece, C., Downes, D. J., Gosden, M., Sharpe, J. A., Sloane-Stanley, J. A., Hughes, J. R., Davies, B., and Higgs, D. R. (2017). Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nat. Cell Biol.*, 19(8):952–961.

Hashimoto, H., Wang, D., Horton, J. R., Zhang, X., Corces, V. G., and Cheng, X. (2017). Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. *Mol. Cell*, 66(5):711–720.

Hattangadi, S. M., Wong, P., Zhang, L., Flygare, J., and Lodish, H. F. (2011). From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood*, 118(24):6258–6268.

Hawse, J. R., Cicek, M., Grygo, S. B., Bruinsma, E. S., Rajamannan, N. M., van Wijnen, A. J., Lian, J. B., Stein, G. S., Oursler, M. J., Subramaniam, M., and Spelsberg, T. C. (2011). TIEG1/KLF10 modulates Runx2 expression and activity in osteoblasts. *PLoS ONE*, 6(4):e19429.

Hay, D., Hughes, J. R., Babbs, C., Davies, J. O., Graham, B. J., Hanssen, L. L., Kassouf, M. T., Oudelaar, A. M., Sharpe, J. A., Suciu, M. C., et al. (2016). Genetic dissection of the [alpha]-globin super-enhancer in vivo. *Nature Genetics*, 48(8):895–903.

He, A., Kong, S. W., Ma, Q., and Pu, W. T. (2011). Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc. Natl. Acad. Sci. U.S.A.*, 108(14):5632–5637.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38(4):576–589.

Hendel, A., Fine, E. J., Bao, G., and Porteus, M. H. (2015). Quantifying on- and off-target genome editing. *Trends Biotechnol.*, 33(2):132–140.

Hertz, G. Z. and Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–577.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, 6(4):283–289.

Higgs, D. R., Engel, J. D., and Stamatoyannopoulos, G. (2012). Thalassaemia. *Lancet*, 379(9813):373–383.

Higgs, D. R., Vernimmen, D., and Wood, B. (2008). Long-range regulation of α-globin gene expression. *Advances in genetics*, 61:143–173.

Hill, J. T., Demarest, B. L., Bisgrove, B. W., Su, Y. C., Smith, M., and Yost, H. J. (2014). Poly peak parser: Method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products. *Dev. Dyn.*, 243(12):1632–1636.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-Andre, V., Sigova, A. A., Hoke, H. A., and Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, 155(4):934–947.

Hnisz, D., Schuijers, J., Lin, C. Y., Weintraub, A. S., Abraham, B. J., Lee, T. I., Bradner, J. E., and Young, R. A. (2015). Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol. Cell*, 58(2):362–370.

Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., and Sharp, P. A. (2017). A phase separation model for transcriptional control. *Cell*, 169(1):13–23.

Hong, J.-W., Hendrix, D. A., and Levine, M. S. (2008). Shadow enhancers as a source of evolutionary novelty. *Science (New York, NY)*, 321(5894):1314.

Horsfield, J. A., Anagnostou, S. H., Hu, J. K., Cho, K. H., Geisler, R., Lieschke, G., Crosier, K. E., and Crosier, P. S. (2007). Cohesin-dependent regulation of Runx genes. *Development*, 134(14):2639–2649.

Hou, C., Li, L., Qin, Z. S., and Corces, V. G. (2012). Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol. Cell*, 48(3):471–484.

Hsieh, T. H., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O. J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, 162(1):108–119.

Huang, D. Y., Kuo, Y. Y., and Chang, Z. F. (2005). GATA-1 mediates auto-regulation of Gfi-1B transcription in K562 cells. *Nucleic Acids Res.*, 33(16):5331–5342.

Huang, J., Liu, X., Li, D., Shao, Z., Cao, H., Zhang, Y., Trompouki, E., Bowman, T. V., Zon, L. I., Yuan, G. C., Orkin, S. H., and Xu, J. (2016). Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. *Dev. Cell*, 36(1):9–23.

Hug, C. B., Grimaldi, A. G., Kruse, K., and Vaquerizas, J. M. (2017). Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell*, 169(2):216–228.

Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., and Higgs, D. R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.*, 46(2):205–212.

Hyle, J., Zhang, Y., Wright, S., Xu, B., Shao, Y., Easton, J., Tian, L., Feng, R., Xu, P., and Li, C. (2019). Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping. *Nucleic Acids Res.*, 47(13):6699–6713.

Iborra, F. J., Pombo, A., Jackson, D. A., and Cook, P. R. (1996). Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. *J. Cell. Sci.*, 109 (Pt 6):1427–1436.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., and Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, 9(10):999–1003.

Irion, S., Nostro, M. C., Kattman, S. J., and Keller, G. M. (2008). Directed differentiation of pluripotent stem cells: from developmental biology to therapeutic applications. *Cold Spring Harb. Symp. Quant. Biol.*, 73:101–110.

Isoda, T., Moore, A. J., He, Z., Chandra, V., Aida, M., Denholtz, M., Piet van Hamburg, J., Fisch, K. M., Chang, A. N., Fahl, S. P., Wiest, D. L., and Murre, C. (2017). Non-coding Transcription Instructs Chromatin Folding and Compartmentalization to Dictate Enhancer-Promoter Communication and T Cell Fate. *Cell*, 171(1):103–119.

Iyer, S., Suresh, S., Guo, D., Daman, K., Chen, J. C. J., Liu, P., Zieger, M., Luk, K., Roscoe, B. P., Mueller, C., King, O. D., Emerson, C. P., and Wolfe, S. A. (2019). Precise therapeutic gene correction by a simple nuclease-induced double-stranded break. *Nature*, 568(7753):561–565.

Jaffredo, T., Gautier, R., Eichmann, A., and Dieterlen-Lièvre, F. (1998). Intraaortic hemopoietic cells are derived from endothelial cells during ontogeny. *Development*, 125(22):4575–4583.

Jeronimo, C., Langelier, M. F., Bataille, A. R., Pascal, J. M., Pugh, B. F., and Robert, F. (2016). Tail and Kinase Modules Differently Regulate Core Mediator Recruitment and Function In Vivo. *Mol. Cell*, 64(3):455–466.

Jeziorska, D. M., Murray, R. J. S., De Gobbi, M., Gaentzsch, R., Garrick, D., Ayyub, H., Chen, T., Li, E., Telenius, J., Lynch, M., Graham, B., Smith, A. J. H., Lund, J. N., Hughes, J. R., Higgs, D. R., and Tufarelli, C. (2017). DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proc. Natl. Acad. Sci. U.S.A.*, 114(36):E7526–E7535.

Jiang, T., Raviram, R., Snetkova, V., Rocha, P. P., Proudhon, C., Badri, S., Bonneau, R., Skok, J. A., and Kluger, Y. (2016). Identification of multi-loci hubs from 4C-seq demonstrates the functional importance of simultaneous interactions. *Nucleic Acids Res.*, 44(18):8714–8725.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821.

Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, 36(16):5221–5231.

Kadonaga, J. T. (2002). The DPE, a core promoter element for transcription by RNA polymerase II. *Exp. Mol. Med.*, 34(4):259–264.

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J., and Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435.

Karin, M. (1995). The regulation of AP-1 activity by mitogen-activated protein kinases. *J. Biol. Chem.*, 270(28):16483–16486.

Kartalaei, P. S., Yamada-Inagawa, T., Vink, C. S., de Pater, E., van der Linden, R., Marks-Bluth, J., van der Sloot, A., van den Hout, M., Yokomizo, T., van Schaick-Solernó, M. L., et al. (2015). Whole-transcriptome analysis of endothelial to hematopoietic stem cell transition reveals a requirement for gpr56 in hsc generation. *The Journal of experimental medicine*, 212(1):93–106.

Kearns, N. A., Pham, H., Tabak, B., Genga, R. M., Silverstein, N. J., Garber, M., and Maehr, R. (2015). Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat. Methods*, 12(5):401–403.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.*, 12(4):656–664.

Kielbasa, S. M. and Vingron, M. (2008). Transcriptional autoregulatory loops are highly conserved in vertebrate evolution. *PLoS ONE*, 3(9):e3210.

Kim, L. K., Esplugues, E., Zorca, C. E., Parisi, F., Kluger, Y., Kim, T. H., Galjart, N. J., and Flavell, R. A. (2014). Oct-1 regulates IL-17 expression by directing interchromosomal associations in conjunction with CTCF in T cells. *Mol. Cell*, 54(1):56–66.

Kim, S. I., Matsumoto, T., Kagawa, H., Nakamura, M., Hirohata, R., Ueno, A., Ohishi, M., Sakuma, T., Soga, T., Yamamoto, T., and Woltjen, K. (2018). Microhomology-assisted scarless genome editing in human iPSCs. *Nat Commun*, 9(1):939.

Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., and Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187.

Kioussis, D., Vanin, E., deLange, T., Flavell, R. A., and Grosveld, F. G. (1983). Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature*, 306(5944):662–666.

Kissa, K. and Herbomel, P. (2010). Blood stem cells emerge from aortic endothelium by a novel type of cell transition. *Nature*, 464(7285):112–115.

Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T. K., Zacarias-Cabeza, J., Spicuglia, S., de la Chapelle, A. L., Heidemann, M., Hintermair, C., Eick, D., Gut, I., Ferrier, P., and Andrau, J. C. (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.*, 18(8):956–963.

Koike-Yusa, H., Li, Y., Tan, E. P., Velasco-Herrera, M. d. e. l. C., and Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.*, 32(3):267–273.

Kojic, A., Cuadrado, A., De Koninck, M., Gimenez-Llorente, D., Rodriguez-Corsino, M., Gomez-Lopez, G., Le Dily, F., Marti-Renom, M. A., and Losada, A. (2018). Distinct roles of cohesin-SA1 and cohesin-SA2 in 3D chromosome organization. *Nat. Struct. Mol. Biol.*, 25(6):496–504.

Korkmaz, G., Lopes, R., Ugalde, A. P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R., and Agami, R. (2016). Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.*, 34(2):192–198.

Kosicki, M., Tomberg, K., and Bradley, A. (2018). Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.*, 36(8):765–771.

Kowalczyk, M. S., Hughes, J. R., Garrick, D., Lynch, M. D., Sharpe, J. A., Sloane-Stanley, J. A., McGowan, S. J., De Gobbi, M., Hosseini, M., Vernimmen, D., Brown, J. M., Gray, N. E., Collavin, L., Gibbons, R. J., Flint, J., Taylor, S., Buckle, V. J., Milne, T. A., Wood, W. G., and Higgs, D. R. (2012). Intragenic enhancers act as alternative promoters. *Mol. Cell*, 45(4):447–458.

Kraft, K., Magg, A., Heinrich, V., Riemenschneider, C., Schopflin, R., Markowski, J., Ibrahim, D. M., Acuna-Hidalgo, R., Despang, A., Andrey, G., Wittler, L., Timmermann, B., Vingron, M., and Mundlos, S. (2019). Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. *Nat. Cell Biol.*, 21(3):305–310.

Krijger, P. H. L. and de Laat, W. (2017). Can We Just Say: Transcription Second? *Cell*, 169(2):184–185.

Kumano, K., Chiba, S., Kunisato, A., Sata, M., Saito, T., Nakagami-Yamaguchi, E., Yamaguchi, T., Masuda, S., Shimizu, K., Takahashi, T., et al. (2003). Notch1 but not notch2 is essential for generating hematopoietic stem cells from endothelial cells. *Immunity*, 18(5):699–711.

Kvon, E. Z., Kamneva, O. K., Melo, U. S., Barozzi, I., Osterwalder, M., Mannion, B. J., Tissieres, V., Pickle, C. S., Plajzer-Frick, I., Lee, E. A., Kato, M., Garvin, T. H., Akiyama, J. A., Afzal, V., Lopez-Rios, J., Rubin, E. M., Dickel, D. E., Pennacchio, L. A., and Visel, A. (2016). Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell*, 167(3):633–642.

Kwiatkowski, N., Zhang, T., Rahl, P. B., Abraham, B. J., Reddy, J., Ficarro, S. B., Dastur, A., Amzallag, A., Ramaswamy, S., Tesar, B., et al. (2014). Targeting transcription regulation in cancer with a covalent cdk7 inhibitor. *Nature*, 511(7511):616.

Lacaud, G., Gore, L., Kennedy, M., Kouskoff, V., Kingsley, P., Hogan, C., Carlsson, L., Speck, N., Palis, J., and Keller, G. (2002). Runx1 is essential for hematopoietic commitment at the hemangioblast stage of development in vitro. *Blood*, 100(2):458–466.

Lacaud, G., Keller, G., and Kouskoff, V. (2004). Tracking mesoderm formation and specification to the hemangioblast in vitro. *Trends in Cardiovascular Medicine*, 14(8):314 – 317.

Lam, D. D., de Souza, F. S., Nasif, S., Yamashita, M., Lopez-Leal, R., Otero-Corchon, V., Meece, K., Sampath, H., Mercer, A. J., Wardlaw, S. L., Rubinstein, M., and Low, M. J. (2015). Partially redundant enhancers cooperatively maintain Mammalian pomc expression above a critical functional threshold. *PLoS Genet.*, 11(2):e1004935.

Lancrin, C., Mazan, M., Stefanska, M., Patel, R., Lichtinger, M., Costa, G., Vargel, O., Wilson, N. K., Moroy, T., Bonifer, C., Gottgens, B., Kouskoff, V., and Lacaud, G. (2012). GFI1 and GFI1B control the loss of endothelial identity of hemogenic endothelium during hematopoietic commitment. *Blood*, 120(2):314–322.

Lancrin, C., Sroczynska, P., Stephenson, C., Allen, T., Kouskoff, V., and Lacaud, G. (2009). The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. *Nature*, 457(7231):892–895.

Ledran, M. H., Krassowska, A., Armstrong, L., Dimmick, I., Renstrom, J., Lang, R., Yung, S., Santibanez-Coref, M., Dzierzak, E., Stojkovic, M., Oostendorp, R. A., Forrester, L., and Lako, M. (2008). Efficient hematopoietic differentiation of human embryonic stem cells on stromal cells derived from hematopoietic niches. *Cell Stem Cell*, 3(1):85–98.

Lee, D., Park, C., Lee, H., Lugus, J. J., Kim, S. H., Arentson, E., Chung, Y. S., Gomez, G., Kyba, M., Lin, S., et al. (2008). Er71 acts downstream of bmp, notch, and wnt signaling in blood and vessel progenitor specification. *Cell stem cell*, 2(5):497–507.

Lee, J., K. I. D. R. K. and Dean, A. (2017). The ldb1 complex co-opts ctcf for erythroid lineage-specific long-range enhancer interactions. *Cell Rep*, 19(12):2490–2502.

Lee, G. R., S. C. G. and Flavell, R. A. (2005). Hypersensitive site 7 of the th2 locus control region is essential for expressing th2 cytokine genes and for long-range intrachromosomal interactions. *Nat Immunol*, 6(1):42–8.

Lee, T. I. and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, 13(4):233.

Lettice, L. A., Daniels, S., Sweeney, E., Venkataraman, S., Devenney, P. S., Gautier, P., Morrison, H., Fantes, J., Hill, R. E., and FitzPatrick, D. R. (2011). Enhancer-adoption as a mechanism of human developmental disease. *Hum. Mutat.*, 32(12):1492–1499.

Levanon, D., Glusman, G., Bangsow, T., Ben-Asher, E., Male, D. A., Avidan, N., Bangsow, C., Hattori, M., Taylor, T. D., Taudien, S., Blechschmidt, K., Shimizu, N., Rosenthal, A., Sakaki, Y., Lancet, D., and Groner, Y. (2001). Architecture and anatomy of the genomic locus encoding the human leukemia-associated transcription factor RUNX1/AML1. *Gene*, 262(1-2):23–33.

Levanon, D. and Groner, Y. (2004). Structure and regulated expression of mammalian RUNX genes. *Oncogene*, 23(24):4211–4219.

Li, L., Rispoli, R., Patient, R., Ciau-Uitz, A., and Porcher, C. (2019). Etv6 activates vegfa expression through positive and negative transcriptional regulatory networks in Xenopus embryos. *Nat Commun*, 10(1):1083.

Li, P., Lahvic, J. L., Binder, V., Pugach, E. K., Riley, E. B., Tamplin, O. J., Panigrahy, D., Bowman, T. V., Barrett, F. G., Heffner, G. C., McKinney-Freeman, S., Schlaeger, T. M., Daley, G. Q., Zeldin, D. C., and Zon, L. I. (2015). Epoxyeicosatrienoic acids enhance embryonic haematopoiesis and adult marrow engraftment. *Nature*, 523(7561):468–471.

Li, Q., Peterson, K. R., Fang, X., and Stamatoyannopoulos, G. (2002). Locus control regions. *Blood*, 100(9):3077–3086.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A. Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., Oh, S., Kim, H. S., Glass, C. K., and Rosenfeld, M. G. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, 498(7455):516–520.

Li, X. and Noll, M. (1994). Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the Drosophila embryo. *EMBO J.*, 13(2):400–406.

Li, Y., Esain, V., Teng, L., Xu, J., Kwan, W., Frost, I. M., Yzaguirre, A. D., Cai, X., Cortes, M., Maijenburg, M. W., et al. (2014a). Inflammatory signaling regulates embryonic hematopoietic stem and progenitor cell production. *Genes & development*, 28(23):2597–2612.

Li, Y., Rivera, C. M., Ishii, H., Jin, F., Selvaraj, S., Lee, A. Y., Dixon, J. R., and Ren, B. (2014b). CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS ONE*, 9(12):e114485.

Lie-A-Ling, M., Marinopoulou, E., Lilly, A. J., Challinor, M., Patel, R., Lancrin, C., Kouskoff, V., and Lacaud, G. (2018). Regulation of RUNX1 dosage is crucial for efficient blood formation from hemogenic endothelium. *Development*, 145(5).

Lieber, M. R. (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem.*, 79:181–211.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293.

Lim, B., Heist, T., Levine, M., and Fukaya, T. (2018). Visualization of Transvection in Living Drosophila Embryos. *Mol. Cell*, 70(2):287–296.

Link, N., Kurtz, P., O'Neal, M., Garcia-Hughes, G., and Abrams, J. M. (2013). A p53 enhancer region regulates target genes through chromatin conformations in cis and in trans. *Genes & development*, 27(22):2433–2438.

Lis, R., Karrasch, C. C., Poulos, M. G., Kunar, B., Redmond, D., Duran, J. G. B., Badwe, C. R., Schachterle, W., Ginsberg, M., Xiang, J., et al. (2017). Conversion of adult endothelium to immunocompetent haematopoietic stem cells. *Nature*, 545(7655):439.

Liu, W., M. Q. W. K. L. W. O. K. Z. J. A. A. and Rosenfeld, M. G. (2013). Brd4 and jmjd6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell*, 155(7):1581–1595.

Liu, Z., Scannell, D. R., Eisen, M. B., and Tjian, R. (2011). Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell*, 146(5):720–731.

Lo, K. W., Kan, H. M., Ashe, K. M., and Laurencin, C. T. (2012). The small molecule PKA-specific cyclic AMP analogue as an inducer of osteoblast-like cells differentiation and mineralization. *J Tissue Eng Regen Med*, 6(1):40–48.

Lobanenkov, V. V., Nicolas, R. H., Adler, V. V., Paterson, H., Klenova, E. M., Polotskaja, A. V., and Goodwin, G. H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene*, 5(12):1743–1753.

Lobanenkov, V. V., Nicolas, R. H., Plumb, M. A., Wright, C. A., and Goodwin, G. H. (1986). Sequence-specific DNA-binding proteins which interact with (G + C)-rich sequences flanking the chicken c-myc gene. *Eur. J. Biochem.*, 159(1):181–188.

Logan, C. V., Cossins, J., Rodriguez Cruz, P. M., Parry, D. A., Maxwell, S., Martinez-Martinez, P., Riepsaame, J., Abdelhamed, Z. A., Lake, A. V., Moran, M., Robb, S., Chow, G., Sewry, C., Hopkins, P. M., Sheridan, E., Jayawant, S., Palace, J., Johnson, C. A., and Beeson, D. (2015). Congenital Myasthenic Syndrome Type 19 Is Caused by Mutations in COL13A1, Encoding the Atypical Non-fibrillar Collagen Type XIII Î±1 Chain. *Am. J. Hum. Genet.*, 97(6):878–885.

Lomvardas, S., Barnea, G., Pisapia, D. J., Mendelsohn, M., Kirkland, J., and Axel, R. (2006). Interchromosomal interactions and olfactory receptor choice. *Cell*, 126(2):403–413.

Long, H. K., Prescott, S. L., and Wysocka, J. (2016). Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell*, 167(5):1170–1187.

Lotem, J., Levanon, D., Negreanu, V., Bauer, O., Hantisteanu, S., Dicken, J., and Groner, Y. (2017). Runx3 in Immunity, Inflammation and Cancer. *Adv. Exp. Med. Biol.*, 962:369–393.

Loukinov, D. I., Pugacheva, E., Vatolin, S., Pack, S. D., Moon, H., Chernukhin, I., Mannan, P., Larsson, E., Kanduri, C., Vostrov, A. A., Cui, H., Niemitz, E. L., Rasko, J. E., Docquier, F. M., Kistler, M., Breen, J. J., Zhuang, Z., Quitschke, W. W., Renkawitz, R., Klenova, E. M., Feinberg, A. P., Ohlsson, R., Morse, H. C., and Lobanenkov, V. V. (2002). BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma. *Proc. Natl. Acad. Sci. U.S.A.*, 99(10):6806–6811.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550.

Loven, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., Bradner, J. E., Lee, T. I., and Young, R. A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, 153(2):320–334.

Lukoseviciute, M., Gavriouchkina, D., Williams, R. M., Hochgreb-Hagele, T., Senanayake, U., Chong-Morrison, V., Thongjuea, S., Repapi, E., Mead, A., and Sauka-Spengler, T. (2018). From Pioneer to Repressor: Bimodal foxd3 Activity Dynamically Remodels Neural Crest Regulatory Landscape In Vivo. *Dev. Cell*, 47(5):608–628.

Lupianez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., and Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025.

Ma, J., Tan, C., Gao, X., Fulbright, R. M., Roberts, J. W., and Wang, M. D. (2019). Transcription factor regulation of RNA polymerase's torque generation capacity. *Proc. Natl. Acad. Sci. U.S.A.*, 116(7):2583–2588.

Ma, Y., Shen, B., Zhang, X., Lu, Y., Chen, W., Ma, J., Huang, X., and Zhang, L. (2014). Heritable multiplex genetic engineering in rats using CRISPR/Cas9. *PLoS ONE*, 9(3):e89413.

Magoc, T. and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963.

Marinic, M., A. T. R. S. and Spitz, F. (2013). An integrated holo-enhancer unit defines tissue and gene specificity of the fgf8 regulatory landscape. *Dev Cell*, 24(5):530–42.

Marsman, J., O'Neill, A. C., Kao, B. R.-Y., Rhodes, J. M., Meier, M., Antony, J., Mönnich, M., and Horsfield, J. A. (2014). Cohesin and ctcf differentially regulate spatiotemporal runx1 expression during zebrafish development. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1839(1):50–61.

Marsman, J., Thomas, A., Osato, M., O'Sullivan, J. M., and Horsfield, J. A. (2017). A DNA Contact Map for the Mouse Runx1 Gene Identifies Novel Haematopoietic Enhancers. *Sci Rep*, 7(1):13347.

Martinez, M., Hinojosa, M., Trombly, D., Morin, V., Stein, J., Stein, G., Javed, A., and Gutierrez, S. E. (2016). Transcriptional Auto-Regulation of RUNX1 P1 Promoter. *PLoS ONE*, 11(2):e0149119.

Maston, G. A., Landt, S. G., Snyder, M., and Green, M. R. (2012). Characterization of enhancer function from genome-wide analyses. *Annual review of genomics and human genetics*, 13:29–57.

Maurano, M. T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., and Stamatoyannopoulos, J. A. (2015). Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.*, 47(12):1393–1401.

McGrath, K. E., F. J. M. F. K. H. B. J. R. C. S. J. C. S. C. K. P. D. K. A. D. and Palis, J. (2015). Distinct sources of hematopoietic progenitors emerge before hscs and provide functional blood cells in the mammalian embryo. *Cell Rep*, 11(12):1892–904.

McVey, M. and Lee, S. E. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet.*, 24(11):529–538.

Medina, M. A., Ugarte, G. D., Vargas, M. F., Avila, M. E., Necuñir, D., Elorza, A. A., Gutiérrez, S. E., and De Ferrari, G. V. (2016). Alternative runx1 promoter regulation by wnt/β-catenin signaling in leukemia cells and human hematopoietic progenitors. *Journal of cellular physiology*.

Medvinsky, A. and Dzierzak, E. (1996). Definitive hematopoiesis is autonomously initiated by the agm region. *Cell*, 86(6):897–906.

Merkenschlager, M. and Odom, D. T. (2013). CTCF and cohesin: linking gene regulatory elements with their targets. *Cell*, 152(6):1285–1297.

Mevel, R., Draper, J. E., Lie-A-Ling, M., Kouskoff, V., and Lacaud, G. (2019). RUNX transcription factors: orchestrators of development. *Development*, 146(17).

Mianne, J., Codner, G. F., Caulder, A., Fell, R., Hutchison, M., King, R., Stewart, M. E., Wells, S., and Teboul, L. (2017). Analysing the outcome of CRISPR-aided genome editing in embryos: Screening, genotyping and quality control. *Methods*, 121-122:68–76.

Mill, C. P., Fiskus, W., DiNardo, C. D., Qian, Y., Raina, K., Rajapakshe, K., Perera, D., Coarfa, C., Kadia, T. M., Khoury, J. D., et al. (2019). Runx1 targeted therapy for aml expressing somatic or germline mutation in runx1. *Blood*, pages blood–2018893982.

Miyoshi, H., Ohira, M., Shimizu, K., Mitani, K., Hirai, H., Imai, T., Yokoyama, K., Soeda, E., and Ohki, M. (1995). Alternative splicing and genomic structure of the AML1 gene involved in acute myeloid leukemia. *Nucleic Acids Res.*, 23(14):2762–2769.

Molyneux, E. M., Rochford, R., Griffin, B., Newton, R., Jackson, G., Menon, G., Harrison, C. J., Israels, T., and Bailey, S. (2012). Burkitt's lymphoma. *Lancet*, 379(9822):1234–1244.

Monahan, K., Horta, A., and Lomvardas, S. (2019). LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. *Nature*, 565(7740):448–453.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F., and Duboule, D. (2011). A regulatory archipelago controls Hox genes transcription in digits. *Cell*, 147(5):1132–1145.

Moore, B. L., Aitken, S., and Semple, C. A. (2015). Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. *Genome Biol.*, 16:110.

Moore, M. A. and Metcalf, D. (1970). Ontogeny of the haemopoietic system: yolk sac origin of in vivo and in vitro colony forming cells in the developing mouse embryo. *British journal of haematology*, 18(3):279–296.

Moreau, P., Hen, R., Wasylyk, B., Everett, R., Gaub, M. P., and Chambon, P. (1981). The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res.*, 9(22):6047–6068.

Muerdter, F., Boryn, M., Woodfin, A. R., Neumayr, C., Rath, M., Zabidi, M. A., Pagani, M., Haberle, V., Kazmar, T., Catarino, R. R., Schernhuber, K., Arnold, C. D., and Stark, A. (2018). Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods*, 15(2):141–149.

Müller, A. M., Medvinsky, A., Strouboulis, J., Grosveld, F., and Dzierzakt, E. (1994). Development of hematopoietic stem cell activity in the mouse embryo. *Immunity*, 1(4):291–301.

Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., and Chang, H. Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, 13(11):919–922.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64.

Nasmyth, K. (2001). Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annu Rev Genet*, 35:673–745.

Nasmyth, K. and Haering, C. H. (2005). The structure and function of smc and kleisin complexes. *Annu Rev Biochem*, 74:595–648.

Nativio, R., Wendt, K. S., Ito, Y., Huddleston, J. E., Uribe-Lewis, S., Woodfine, K., Krueger, C., Reik, W., Peters, J. M., and Murrell, A. (2009). Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. *PLoS Genet.*, 5(11):e1000739.

Navarro-Montero, O., Ayllon, V., Lamolda, M., Lopez-Onieva, L., Montes, R., Bueno, C., Ng, E., Guerrero-Carreno, X., Romero, T., Romero-Moya, D., Stanley, E., Elefanty, A., Ramos-Mejia, V., Menendez, P., and Real, P. J. (2017). RUNX1c Regulates Hematopoietic Differentiation of Human Pluripotent Stem Cells Possibly in Cooperation with Proinflammatory Signaling. *Stem Cells*, 35(11):2253–2266.

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90.

Neuberger, M. S. (1983). Expression and regulation of immunoglobulin heavy chain gene transfected into lymphoid cells. *EMBO J*, 2(8):1373–8.

Ng, C. E. L., Yokomizo, T., Yamashita, N., Cirovic, B., Jin, H., Wen, Z., Ito, Y., and Osato, M. (2010). A runx1 intronic enhancer marks hemogenic endothelial cells and hematopoietic stem cells. *Stem cells*, 28(10):1869–1881.

Nishida, H., Suzuki, T., Ookawa, H., Tomaru, Y., and Hayashizaki, Y. (2005). Comparative analysis of expression of histone H2a genes in mouse. *BMC Genomics*, 6:108.

Noordermeer, D. and de Laat, W. (2008). Joining the loops: β-globin gene regulation. *IUBMB life*, 60(12):824–833.

Nora, E. P., Goloborodko, A., Valton, A. L., Gibcus, J. H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L. A., and Bruneau, B. G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, 169(5):930–944.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Bluthgen, N., Dekker, J., and Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385.

North, T., Gu, T.-L., Stacy, T., Wang, Q., Howard, L., Binder, M., Marín-Padilla, M., and Speck, N. A. (1999). Cbfa2 is required for the formation of intra-aortic hematopoietic clusters. *Development*, 126(11):2563–2575.

Nottingham, W. T., Jarratt, A., Burgess, M., Speck, C. L., Cheng, J.-F., Prabhakar, S., Rubin, E. M., Li, P.-S., Sloane-Stanley, J., Kong-a San, J., et al. (2007). Runx1-mediated hematopoietic stem-cell emergence is controlled by a gata/ets/scl-regulated enhancer. *Blood*, 110(13):4188–4197.

O'Byrne, S., Elliott, N., Rice, S., Buck, G., Fordham, N., Garnett, C., Godfrey, L., Crump, N. T., Wright, G., Inglott, S., Hua, P., Psaila, B., Povinelli, B., Knapp, D. J. H. F., Agraz-Doblas, A., Bueno, C., Varela, I., Bennett, P., Koohy, H., Watt, S. M., Karadimitris, A., Mead, A. J., Ancliff, P., Vyas, P., Menendez, P., Milne, T. A., Roberts, I., and Roy, A. (2019). Discovery of a CD10 negative B-progenitor in human fetal life identifies unique ontogeny-related developmental programs. *Blood*.

Oh, J., Sanders, I. F., Chen, E. Z., Li, H., Tobias, J. W., Isett, R. B., Penubarthi, S., Sun, H., Baldwin, D. A., and Fraser, N. W. (2015). Genome wide nucleosome mapping for HSV-1 shows nucleosomes are deposited at preferred positions during lytic infection. *PLoS ONE*, 10(2):e0117471.

Okada, H., Watanabe, T., Niki, M., Takano, H., Chiba, N., Yanai, N., Tani, K., Hibino, H., Asano, S., Mucenski, M. L., et al. (1998). Aml1 (-/-) embryos do not express certain hematopoiesis-related gene transcripts including those of the pu. 1 gene. *Oncogene*, 17(2287):2293.

Okuda, T., Van Deursen, J., Hiebert, S. W., Grosveld, G., and Downing, J. R. (1996). Aml1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell*, 84(2):321–330.

Ong, C. T. and Corces, V. G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet*, 12(4):283–93.

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., Lee, E. A., Zhu, Y., Plajzer-Frick, I., Pickle, C. S., et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, 554(7691):239.

Ottersbach, K. (2019). Endothelial-to-haematopoietic transition: an update on the process of making blood. *Biochem. Soc. Trans.*, 47(2):591–601.

Oudelaar, A. M., Beagrie, R. A., Gosden, M., de Ornellas, S., Georgiades, E., Kerry, J., Hidalgo, D., Carrelha, J., Shivalingam, A., El-Sagheer, A. H., Telenius, J. M., Brown, T., Buckle, V. J., Socolovsky, M., Higgs, D. R., and Hughes, J. R. (2019). Dissection of the 4d chromatin structure of the a-globin locus through in vivo erythroid differentiation with extreme spatial and temporal resolution. *bioRxiv*.

Oudelaar, A. M., Davies, J. O. J., Hanssen, L. L. P., Telenius, J. M., Schwessinger, R., Liu, Y., Brown, J. M., Downes, D. J., Chiariello, A. M., Bianco, S., Nicodemi, M., Buckle, V. J., Dekker, J., Higgs, D. R., and Hughes, J. R. (2018). Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. *Nat. Genet.*, 50(12):1744–1751.

Oudelaar, A. M., H. L. L. P. H. R. C. K. M. T. H. J. R. and Higgs, D. R. (2017). Between form and function: the complexity of genome folding. *Hum Mol Genet*, 26(R2):R208–R215.

Owens, D. D. G., Caulder, A., Frontera, V., Harman, J. R., Allan, A. J., Bucakci, A., Greder, L., Codner, G. F., Hublitz, P., McHugh, P. J., Teboul, L., and de Bruijn, M. F. T. R. (2019). Microhomologies are prevalent at Cas9-induced larger deletions. *Nucleic Acids Research*, 47(14):7402–7417.

Palis, J., Robertson, S., Kennedy, M., Wall, C., and Keller, G. (1999). Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse. *Development*, 126(22):5073–5084.

Palstra, R. J. and Grosveld, F. (2012). Transcription factor binding at enhancers: shaping a genomic regulatory landscape in flux. *Front Genet*, 3:195.

Palstra, R. J., Simonis, M., Klous, P., Brasset, E., Eijkelkamp, B., and de Laat, W. (2008). Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. *PLoS ONE*, 3(2):e1661.

Papadopoulos, D. K., Skouloudaki, K., Engstr?m, Y., Terenius, L., Rigler, R., Zechner, C., Vukojevi?, V., and Tomancak, P. (2019). Control of Hox transcription factor concentration and cell-to-cell variability by an auto-regulatory switch. *Development*, 146(12).

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., Cobb, B. S., Yokomori, K., Dillon, N., Aragon, L., Fisher, A. G., and Merkenschlager, M. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, 132(3):422–433.

Parikh, B. A., Beckman, D. L., Patel, S. J., White, J. M., and Yokoyama, W. M. (2015). Detailed phenotypic and molecular analyses of genetically modified mice generated by CRISPR-Cas9-mediated editing. *PLoS ONE*, 10(1):e0116484.

Parker, S. C., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., van Bueren, K. L., Chines, P. S., Narisu, N., Black, B. L., Visel, A., Pennacchio, L. A., Collins, F. S., Becker, J., Benjamin, B., Blakesley, R., Bouffard, G., Brooks, S., Coleman, H., Dekhtyar, M., Gregory, M., Guan, X., Gupta, J., Han, J., Hargrove, A., Ho, S. L., Johnson, T., Legaspi, R., Lovett, S., Maduro, Q., Masiello, C., Maskeri, B., McDowell, J., Montemayor, C., Mullikin, J., Park, M., Riebow, N., Schandler, K., Schmidt, B., Sison, C., Stantripop, M., Thomas, J., Thomas, P., Vemulapalli, M., and Young, A. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U.S.A.*, 110(44):17921–17926.

Parry, T. J., Theisen, J. W., Hsu, J.-Y., Wang, Y.-L., Corcoran, D. L., Eustice, M., Ohler, U., and Kadonaga, J. T. (2010). The tct motif, a key component of an rna polymerase ii transcription system for the translational machinery. *Genes & development*, 24(18):2013–2018.

Pearson, S., Cuvertino, S., Fleury, M., Lacaud, G., and Kouskoff, V. (2015). In vivo repopulating activity emerges at the onset of hematopoietic specification during embryonic stem cell differentiation. *Stem Cell Reports*, 4(3):431–444.

Perry, M. W., Boettiger, A. N., Bothma, J. P., and Levine, M. (2010). Shadow enhancers foster robustness of Drosophila gastrulation. *Curr. Biol.*, 20(17):1562–1567.

Petrenko, N., J. Y. W. K. H. and Struhl, K. (2016). Mediator undergoes a compositional change during transcriptional activation. *Mol Cell*, 64(3):443–454.

Phillips, J. E. and Corces, V. G. (2009). Ctcf: master weaver of the genome. *Cell*, 137(7):1194–211.

Phillips-Cremins, J. E., Sauria, M. E., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S., Ong, C. T., Hookway, T. A., Guo, C., Sun, Y., Bland, M. J., Wagstaff, W., Dalton, S., McDevitt, T. C., Sen, R., Dekker, J., Taylor, J., and Corces, V. G. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, 153(6):1281–1295.

Pimanda, J. E., Donaldson, I. J., de Bruijn, M. F., Kinston, S., Knezevic, K., Huckle, L., Piltz, S., Landry, J.-R., Green, A. R., Tannahill, D., et al. (2007). The scl transcriptional network and bmp signaling pathway interact to regulate runx1 activity. *Proceedings of the National Academy of Sciences*, 104(3):840–845.

Pinello, L., Canver, M. C., Hoban, M. D., Orkin, S. H., Kohn, D. B., Bauer, D. E., and Yuan, G. C. (2016). Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat. Biotechnol.*, 34(7):695–697.

Pinto do O, P., Kolterud, A., and Carlsson, L. (1998a). Expression of the LIM-homeobox gene LH2 generates immortalized steel factor-dependent multipotent hematopoietic precursors. *EMBO J.*, 17(19):5744–5756.

Pinto do O, P., Kolterud, A., and Carlsson, L. (1998b). Expression of the LIM-homeobox gene LH2 generates immortalized steel factor-dependent multipotent hematopoietic precursors. *EMBO J.*, 17(19):5744–5756.

Piper, J., Assi, S. A., Cauchy, P., Ladroue, C., Cockerill, P. N., Bonifer, C., and Ott, S. (2015). Wellington-bootstrap: differential DNase-seq footprinting identifies cell-type determining transcription factors. *BMC Genomics*, 16:1000.

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, 21(3):447–455.

Plank, J. L. and Dean, A. (2014). Enhancer function: mechanistic and genome-wide insights come together. *Molecular cell*, 55(1):5–14.

Ponjavic, J., Lenhard, B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Sandelin, A. (2006). Transcriptional and structural impact of tata-initiation site spacing in mammalian core promoters. *Genome biology*, 7(8):R78.

Pott, S. and Lieb, J. D. (2015). What are super-enhancers? *Nat Genet*, 47(1):8–12.

Pozner, A., Lotem, J., Xiao, C., Goldenberg, D., Brenner, O., Negreanu, V., Levanon, D., and Groner, Y. (2007). Developmentally regulated promoter-switch transcriptionally controls Runx1 function during embryonic hematopoiesis. *BMC Dev. Biol.*, 7:84.

Preston, J. C., Hileman, L. C., and Cubas, P. (2011). Reduce, reuse, and recycle: developmental evolution of trait diversification. *Am. J. Bot.*, 98(3):397–403.

Proudhon, C., Snetkova, V., Raviram, R., Lobry, C., Badri, S., Jiang, T., Hao, B., Trimarchi, T., Kluger, Y., Aifantis, I., Bonneau, R., and Skok, J. A. (2016). Active and Inactive Enhancers Cooperate to Exert Localized and Long-Range Control of Gene Regulation. *Cell Rep*, 15(10):2159–2169.

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., and Lim, W. A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, 152(5):1173–1183.

Rahl, P. B., Lin, C. Y., Seila, A. C., Flynn, R. A., McCuine, S., Burge, C. B., Sharp, P. A., and Young, R. A. (2010). c-Myc regulates transcriptional pause release. *Cell*, 141(3):432–445.

Ramirez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Gruning, B. A., Villaveces, J., Habermann, B., Akhtar, A., and Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*, 9(1):189.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nat Protoc*, 8(11):2281–2308.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., et al. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680.

Rao, S. S. P., Huang, S. C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K. R., Sanborn, A. L., Johnstone, S. E., Bascom, G. D., Bochkov, I. D., Huang, X., Shamim, M. S., Shin, J., Turner, D., Ye, Z., Omer, A. D., Robinson, J. T., Schlick, T., Bernstein, B. E., Casellas, R., Lander, E. S., and Aiden, E. L. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell*, 171(2):305–320.

Reik, A., Telling, A., Zitnik, G., Cimbora, D., Epner, E., and Groudine, M. (1998). The locus control region is necessary for gene expression in the human beta-globin locus but not the maintenance of an open chromatin structure in erythroid cells. *Mol. Cell. Biol.*, 18(10):5992–6000.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425.

Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., Larson, D. R., and Zhao, K. (2017). CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Mol. Cell*, 67(6):1049–1058.

Ren, X., Siegel, R., Kim, U., and Roeder, R. G. (2011). Direct interactions of OCA-B and TFII-I regulate immunoglobulin heavy-chain gene transcription by facilitating enhancer-promoter communication. *Mol. Cell*, 42(3):342–355.

Rieger, M. A. and Schroeder, T. (2012). Hematopoiesis. *Cold Spring Harb Perspect Biol*, 4(12).

Riu, E., Chen, Z. Y., Xu, H., He, C. Y., and Kay, M. A. (2007). Histone modifications are associated with the persistence or silencing of vector-mediated transgene expression in vivo. *Mol. Ther.*, 15(7):1348–1355.

Robert-Moreno, À., Espinosa, L., de la Pompa, J. L., and Bigas, A. (2005). Rbpjκ-dependent notch function regulates gata2 and is essential for the formation of intra-embryonic hematopoietic cells. *Development*, 132(5):1117–1126.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4(8):651–657.

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., and Bernstein, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, 33(11):1165–1172.

Rubio, E. D., Reiss, D. J., Welcsh, P. L., Disteche, C. M., Filippova, G. N., Baliga, N. S., Aebersold, R., Ranish, J. A., and Krumm, A. (2008). CTCF physically links cohesin to chromatin. *Proc. Natl. Acad. Sci. U.S.A.*, 105(24):8309–8314.

Ruiz-Velasco, M., Kumar, M., Lai, M. C., Bhat, P., Solis-Pinson, A. B., Reyes, A., Kleinsorg, S., Noh, K. M., Gibson, T. J., and Zaugg, J. B. (2017). CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals. *Cell Syst*, 5(6):628–637.

Ruthenburg, A. J., Li, H., Patel, D. J., and Allis, C. D. (2007). Multivalent engagement of chromatin modifications by linked binding modules. *Nat. Rev. Mol. Cell Biol.*, 8(12):983–994.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T. C., Robins, A. J., Dalton, S., and Gilbert, D. M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.*, 20(6):761–770.

Sabari, B. R., Dall'Agnese, A., Boija, A., Klein, I. A., Coffey, E. L., Shrinivas, K., Abraham, B. J., Hannett, N. M., Zamudio, A. V., Manteiga, J. C., Li, C. H., Guo, Y. E., Day, D. S., Schuijers, J., Vasile, E., Malik, S., Hnisz, D., Lee, T. I., Cisse, I. I., Roeder, R. G., Sharp, P. A., Chakraborty, A. K., and Young, R. A. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, 361(6400).

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development*, 132(4):797–803.

Saint-André, V., Federation, A. J., Lin, C. Y., Abraham, B. J., Reddy, J., Lee, T. I., Bradner, J. E., and Young, R. A. (2016). Models of human core transcriptional regulatory circuitries. *Genome research*, 26(3):385–396.

Sanborn, A. L., Rao, S. S., Huang, S. C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S., and Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U.S.A.*, 112(47):E6456–6465.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32(Database issue):D91–94.

Sander, J. D. and Joung, J. K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.*, 32(4):347–355.

Sanjana, N. E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A., and Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. *Science*, 353(6307):1545–1549.

Sawado, T., Halow, J., Bender, M. A., and Groudine, M. (2003). The beta -globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. *Genes Dev.*, 17(8):1009–1018.

Sawamiphak, S., Kontarakis, Z., and Stainier, D. Y. (2014). Interferon gamma signaling positively regulates hematopoietic stem cell emergence. *Developmental cell*, 31(5):640–653.

Saxena, M., Roman, A. K. S., O'Neill, N. K., Sulahian, R., Jadhav, U., and Shivdasani, R. A. (2017). Transcription factor-dependent 'anti-repressive' mammalian enhancers exclude H3K27me3 from extended genomic domains. *Genes Dev.*, 31(23-24):2391–2404.

Schmidt, D., Schwalie, P. C., Ross-Innes, C. S., Hurtado, A., Brown, G. D., Carroll, J. S., Flicek, P., and Odom, D. T. (2010). A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.*, 20(5):578–588.

Schoenfelder, S. and Fraser, P. (2019). Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.*, 20(8):437–455.

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B., and Cavalli, G. (2007). Genome regulation by polycomb and trithorax proteins. *Cell*, 128(4):735–745.

Schuijers, J., Manteiga, J. C., Weintraub, A. S., Day, D. S., Zamudio, A. V., Hnisz, D., Lee, T. I., and Young, R. A. (2018). Transcriptional dysregulation of myc reveals common enhancer-docking mechanism. *Cell reports*, 23(2):349–360.

Schütte, J., Wang, H., Antoniou, S., Jarratt, A., Wilson, N. K., Riepsaame, J., Calero-Nieto, F. J., Moignard, V., Basilico, S., Kinston, S. J., et al. (2016). An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability. *Elife*, 5:e11469.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N. A., Huber, W., H Haering, C., Mirny, L., and Spitz, F. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678):51–56.

Schwessinger, R., Suciu, M. C., McGowan, S. J., Telenius, J., Taylor, S., Higgs, D. R., and Hughes, J. R. (2017). Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. *Genome Res.*, 27(10):1730–1742.

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*, 16:259.

Sfeir, A. and Symington, L. S. (2015). Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends Biochem. Sci.*, 40(11):701–714.

Shalaby, F., Rossant, J., Yamaguchi, T. P., Gertsenstein, M., Wu, X.-F., Breitman, M. L., and Schuh, A. C. (1995). Failure of blood-island formation and vasculogenesis in flk-1-deficient mice. *Nature*, 376(6535):62–66.

Shaulian, E. and Karin, M. (2002). AP-1 as a regulator of cell life and death. *Nat. Cell Biol.*, 4(5):E131–136.

Shen, M. W., Arbab, M., Hsu, J. Y., Worstell, D., Culbertson, S. J., Krabbe, O., Cassa, C. A., Liu, D. R., Gifford, D. K., and Sherwood, R. I. (2018). Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature*, 563(7733):646–651.

Shin, H. Y., Wang, C., Lee, H. K., Yoo, K. H., Zeng, X., Kuhns, T., Yang, C. M., Mohr, T., Liu, C., and Hennighausen, L. (2017). CRISPR/Cas9 targeting events cause complex deletions and insertions at 17 sites in the mouse genome. *Nat Commun*, 8:15464.

Shin, H. Y., Willi, M., Yoo, K. H., Zeng, X., Wang, C., Metser, G., and Hennighausen, L. (2016). Hierarchy within the mammary stat5-driven wap super-enhancer. *Nature genetics*, 48(8):904.

Shipony, Z., Mukamel, Z., Cohen, N. M., Landan, G., Chomsky, E., Zeliger, S. R., Fried, Y. C., Ainbinder, E., Friedman, N., and Tanay, A. (2014). Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature*, 513(7516):115–119.

Shivdasani, R. A., Mayer, E. L., and Orkin, S. H. (1995). Absence of blood formation in mice lacking the t-cell leukaemia oncoprotein tal-1/scl.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479(7371):74–79.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8):1034–1050.

Sima, J., Chakraborty, A., Dileep, V., Michalski, M., Klein, K. N., Holcomb, N. P., Turner, J. L., Paulsen, M. T., Rivera-Mulia, J. C., Trevilla-Garcia, C., Bartlett, D. A., Zhao, P. A., Washburn, B. K., Nora, E. P., Kraft, K., Mundlos, S., Bruneau, B. G., Ljungman, M., Fraser, P., Ay, F., and Gilbert, D. M. (2019). Identifying cis Elements for Spatiotemporal Control of Mammalian DNA Replication. *Cell*, 176(4):816–830.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.*, 38(11):1348–1354.

Small, S., Kraut, R., Hoey, T., Warrior, R., and Levine, M. (1991). Transcriptional regulation of a pair-rule stripe in Drosophila. *Genes Dev.*, 5(5):827–839.

Souilhol, C., Lendinez, J. G., Rybtsov, S., Murphy, F., Wilson, H., Hills, D., Batsivari, A., Binagui-Casas, A., McGarvey, A. C., MacDonald, H. R., et al. (2016). Developing hscs become notch independent by the end of maturation in the agm region. *Blood*, 128(12):1567–1577.

Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R., and Flavell, R. A. (2005). Interchromosomal associations between alternatively expressed loci. *Nature*, 435(7042):637–645.

Sroczynska, P., Lancrin, C., Kouskoff, V., and Lacaud, G. (2009a). The differential activities of Runx1 promoters define milestones during embryonic hematopoiesis. *Blood*, 114(26):5279–5289.

Sroczynska, P., Lancrin, C., Pearson, S., Kouskoff, V., and Lacaud, G. (2009b). In vitro differentiation of mouse embryonic stem cells as a model of early hematopoietic development. *Methods Mol. Biol.*, 538:317–334.

Stedman, W., Kang, H., Lin, S., Kissil, J. L., Bartolomei, M. S., and Lieberman, P. M. (2008). Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *EMBO J.*, 27(4):654–666.

Sugimura, R., Jha, D. K., Han, A., Soria-Valles, C., Da Rocha, E. L., Lu, Y.-F., Goettel, J. A., Serrao, E., Rowe, R. G., Malleshaiah, M., et al. (2017). Haematopoietic stem and progenitor cells from human pluripotent stem cells. *Nature*, 545(7655):432.

Sullivan, J. C., Sher, D., Eisenstein, M., Shigesada, K., Reitzel, A. M., Marlow, H., Levanon, D., Groner, Y., Finnerty, J. R., and Gat, U. (2008). The evolutionary origin of the Runx/CBFbeta transcription factors–studies of the most basal metazoans. *BMC Evol. Biol.*, 8:228.

Swiers, G., Baumann, C., O'Rourke, J., Giannoulatou, E., Taylor, S., Joshi, A., Moignard, V., Pina, C., Bee, T., Kokkaliaris, K. D., et al. (2013a). Early dynamic fate changes in haemogenic endothelium characterized at the single-cell level. *Nature communications*, 4.

Swiers, G., De Bruijn, M., and Speck, N. A. (2010). Hematopoietic stem cell emergence in the conceptus and the role of runx1. *The International journal of developmental biology*, 54:1151.

Swiers, G., Rode, C., Azzoni, E., and de Bruijn, M. F. (2013b). A short history of hemogenic endothelium. *Blood Cells Mol. Dis.*, 51(4):206–212.

Symington, L. S. and Gautier, J. (2011). Double-strand break end resection and repair pathway choice. *Annu. Rev. Genet.*, 45:247–271.

Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Res.*, 24(3):390–400.

Szabo, P. E., Tang, S. H., Silva, F. J., Tsark, W. M., and Mann, J. R. (2004). Role of CTCF binding sites in the Igf2/H19 imprinting control region. *Mol. Cell. Biol.*, 24(11):4791–4800.

Szabo, Q., Bantignies, F., and Cavalli, G. (2019). Principles of genome folding into topologically associating domains. *Science advances*, 5(4):eaaw1668.

Taft, R. J., Pheasant, M., and Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, 29(3):288–299.

Taheri-Ghahfarokhi, A., Taylor, B. J. M., Nitsch, R., Lundin, A., Cavallo, A. L., Madeyski-Bengtson, K., Karlsson, F., Clausen, M., Hicks, R., Mayr, L. M., Bohlooly-Y, M., and Maresca, M. (2018). Decoding non-random mutational signatures at Cas9 targeted sites. *Nucleic Acids Res.*, 46(16):8417–8434.

Takahashi, A., de Andres, M. C., Hashimoto, K., Itoi, E., Otero, M., Goldring, M. B., and Oreffo, R. O. C. (2017). DNA methylation of the RUNX2 P1 promoter mediates MMP13 transcription in chondrocytes. *Sci Rep*, 7(1):7771.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–872.

Takei, H. and Kobayashi, S. S. (2019). Targeting transcription factors in acute myeloid leukemia. *Int. J. Hematol.*, 109(1):28–34.

Tamm, C., Pijuan Galito, S., and Anneren, C. (2013). A comparative study of protocols for mouse embryonic stem cell culturing. *PLoS ONE*, 8(12):e81156.

Tan, E. P., Li, Y., Velasco-Herrera, M. d. e. l. C., Yusa, K., and Bradley, A. (2015). Off-target assessment of CRISPR-Cas9 guiding RNAs in human iPS and mouse ES cells. *Genesis*, 53(2):225–236.

Telenius, J., Consortium, T. W., and Hughes, J. R. (2018). NGseqBasic - a single-command unix tool for atac-seq, dnasei-seq, cut-and-run, and chip-seq data mapping, high-resolution visualisation, and quality control. *bioRxiv*.

Telfer, J. C. and Rothenberg, E. V. (2001). Expression and function of a stem cell promoter for the murine CBFalpha2 gene: distinct roles and regulation in natural killer and T cell development. *Dev. Biol.*, 229(2):363–382.

Thakore, P. I., D'Ippolito, A. M., Song, L., Safi, A., Shivakumar, N. K., Kabadi, A. M., Reddy, T. E., Crawford, G. E., and Gersbach, C. A. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods*, 12(12):1143–1149.

Thambyrajah, R., Mazan, M., Patel, R., Moignard, V., Stefanska, M., Marinopoulou, E., Li, Y., Lancrin, C., Clapes, T., Möröy, T., et al. (2016). Gfi1 proteins orchestrate the emergence of haematopoietic stem cells through recruitment of lsd1. *Nature cell biology*, 18(1):21–32.

Thomas, M., Burgio, G., Adams, D. J., and Iyer, V. (2019). Collateral damage and CRISPR genome editing. *PLoS Genet.*, 15(3):e1007994.

Tiwari, N., Meyer-Schaller, N., Arnold, P., Antoniadis, H., Pachkov, M., van Nimwegen, E., and Christofori, G. (2013). Klf4 is a transcriptional regulator of genes critical for EMT, including Jnk1 (Mapk8). *PLoS ONE*, 8(2):e57329.

Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol. Cell*, 10(6):1453–1465.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 23(1):137–144.

Tsai, F.-Y., Keller, G., Kuo, F. C., Weiss, M., Chen, J., Rosenblatt, M., Alt, F. W., and Orkin, S. H. (1994). An early haematopoietic defect in mice lacking the transcription factor gata-2. *Nature*, 371(6494):221.

Tsukiji, N., Amano, T., and Shiroishi, T. (2014). A novel regulatory element for shh expression in the lung and gut of mouse embryos. *Mechanisms of development*, 131:127–136.

Uenishi, G., Theisen, D., Lee, J. H., Kumar, A., Raymond, M., Vodyanik, M., Swanson, S., Stewart, R., Thomson, J., and Slukvin, I. (2014). Tenascin C promotes hematoendothelial development and T lymphoid commitment from human pluripotent stem cells in chemically defined conditions. *Stem Cell Reports*, 3(6):1073–1084.

van Arensbergen, J., van Steensel, B., and Bussemaker, H. J. (2014). In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol.*, 24(11):695–702.

Van Bortle, K., Nichols, M. H., Li, L., Ong, C. T., Takenaka, N., Qin, Z. S., and Corces, V. G. (2014). Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol.*, 15(6):R82.

van Overbeek, M., Capurso, D., Carter, M. M., Thompson, M. S., Frias, E., Russ, C., Reece-Hoyes, J. S., Nye, C., Gradia, S., Vidal, B., Zheng, J., Hoffman, G. R., Fuller, C. K., and May, A. P. (2016). DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Mol. Cell*, 63(4):633–646.

Vatolin, S., Abdullaev, Z., Pack, S. D., Flanagan, P. T., Custer, M., Loukinov, D. I., Pugacheva, E., Hong, J. A., Morse, H., Schrump, D. S., Risinger, J. I., Barrett, J. C., and Lobanenkov, V. V. (2005). Conditional expression of the CTCF-paralogous transcriptional factor BORIS in normal cells results in demethylation and derepression of MAGE-A1 and reactivation of other cancer-testis genes. *Cancer Res.*, 65(17):7751–7762.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman,

C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Vian, L., Pekowska, A., Rao, S. S. P., Kieffer-Kwon, K. R., Jung, S., Baranello, L., Huang, S. C., El Khattabi, L., Dose, M., Pruett, N., Sanborn, A. L., Canela, A., Maman, Y., Oksanen, A., Resch, W., Li, X., Lee, B., Kovalchuk, A. L., Tang, Z., Nelson, S., Di Pierro, M., Cheng, R. R., Machol, I., St Hilaire, B. G., Durand, N. C., Shamim, M. S., Stamenova, E. K., Onuchic, J. N., Ruan, Y., Nussenzweig, A., Levens, D., Aiden, E. L., and Casellas, R. (2018). The Energetics and Physiological Impact of Cohesin Extrusion. *Cell*, 173(5):1165–1178.

Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R. S., Stehling-Sun, S., Sabo, P. J., Byron, R., Humbert, R., et al. (2014). Mouse regulatory dna landscapes reveal global principles of cis-regulatory evolution. *Science*, 346(6212):1007–1012.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep*, 10(8):1297–1309.

Vimalraj, S., Arumugam, B., Miranda, P. J., and Selvamurugan, N. (2015). Runx2: Structure, function, and phosphorylation in osteoblast differentiation. *Int. J. Biol. Macromol.*, 78:202–208.

Vincent, B. J., Estrada, J., and DePace, A. H. (2016). The appeasement of Doug: a synthetic approach to enhancer biology. *Integr Biol (Camb)*, 8(4):475–484.

Visel, A., Prabhakar, S., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E. M., and Pennacchio, L. A. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.*, 40(2):158–160.

Visel, A., Taher, L., Girgis, H., May, D., Golonzhka, O., Hoch, R. V., McKinsey, G. L., Pattabiraman, K., Silberberg, S. N., Blow, M. J., Hansen, D. V., Nord, A. S., Akiyama, J. A., Holt,

A., Hosseini, R., Phouanenavong, S., Plajzer-Frick, I., Shoukry, M., Afzal, V., Kaplan, T., Kriegstein, A. R., Rubin, E. M., Ovcharenko, I., Pennacchio, L. A., and Rubenstein, J. L. (2013). A high-resolution enhancer atlas of the developing telencephalon. *Cell*, 152(4):895–908.

Wamstad, J. A., Alexander, J. M., Truty, R. M., Shrikumar, A., Li, F., Eilertson, K. E., Ding, H., Wylie, J. N., Pico, A. R., Capra, J. A., Erwin, G., Kattman, S. J., Keller, G. M., Srivastava, D., Levine, S. S., Pollard, K. S., Holloway, A. K., Boyer, L. A., and Bruneau, B. G. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*, 151(1):206–220.

Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M. U., Ohgi, K. A., Glass, C. K., Rosenfeld, M. G., and Fu, X. D. (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, 474(7351):390–394.

Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., Thurman, R. E., Kaul, R., Myers, R. M., and Stamatoyannopoulos, J. A. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.*, 22(9):1680–1688.

Wang, H., Yang, H., Shivalila, C. S., Dawlaty, M. M., Cheng, A. W., Zhang, F., and Jaenisch, R. (2013). One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell*, 153(4):910–918.

Wang, L. C., Kuo, F., Fujiwara, Y., Gilliland, D. G., Golub, T. R., and Orkin, S. H. (1997). Yolk sac angiogenic defect and intra-embryonic apoptosis in mice lacking the ets-related factor tel. *The EMBO journal*, 16(14):4374–4383.

Wang, N., Miao, H., Li, Y. S., Zhang, P., Haga, J. H., Hu, Y., Young, A., Yuan, S., Nguyen, P., Wu, C. C., and Chien, S. (2006). Shear stress regulation of Krüppel-like factor 2 expression is flow pattern-specific. *Biochem. Biophys. Res. Commun.*, 341(4):1244–1251.

Wang, Q., Stacy, T., Binder, M., Marin-Padilla, M., Sharpe, A. H., and Speck, N. A. (1996). Disruption of the cbfa2 gene causes necrosis and hemorrhaging in the central nervous system and blocks definitive hematopoiesis. *Proceedings of the National Academy of Sciences*, 93(8):3444–3449.

Wang, Y., Yang, C., Gu, Q., Sims, M., Gu, W., Pfeffer, L. M., and Yue, J. (2015). KLF4 Promotes Angiogenesis by Activating VEGF Signaling in Human Retinal Microvascular Endothelial Cells. *PLoS ONE*, 10(6):e0130341.

Warren, A. J., Colledge, W. H., Carlton, M. B., Evans, M. J., Smith, A. J., and Rabbitts, T. H. (1994). The oncogenic cysteine-rich lim domain protein rbtn2 is essential for erythroid development. *Cell*, 78(1):45–57.

Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 5(4):276–287.

Weatherall, D. J. (2001). Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. *Nat. Rev. Genet.*, 2(4):245–255.

Weintraub, A. S., Li, C. H., Zamudio, A. V., Sigova, A. A., Hannett, N. M., Day, D. S., Abraham, B. J., Cohen, M. A., Nabet, B., Buckley, D. L., Guo, Y. E., Hnisz, D., Jaenisch, R., Bradner, J. E., Gray, N. S., and Young, R. A. (2017). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*, 171(7):1573–1588.

Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., Yahata, K., Imamoto, F., Aburatani, H., Nakao, M., Imamoto, N., Maeshima, K., Shirahige, K., and Peters, J. M. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, 451(7180):796–801.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319.

Wijchers, P. J., Krijger, P. H. L., Geeven, G., Zhu, Y., Denker, A., Verstegen, M. J. A. M., Valdes-Quezada, C., Vermeulen, C., Janssen, M., Teunissen, H., Anink-Groenen, L. C. M., Verschure, P. J., and de Laat, W. (2016). Cause and Consequence of Tethering a SubTAD to Different Nuclear Compartments. *Mol. Cell*, 61(3):461–473.

Wilson, N. K., Foster, S. D., Wang, X., Knezevic, K., Schütte, J., Kaimakis, P., Chilarska, P. M., Kinston, S., Ouwehand, W. H., Dzierzak, E., et al. (2010a). Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell stem cell*, 7(4):532–544.

Wilson, N. K., Schoenfelder, S., Hannah, R., Castillo, M. S., Schütte, J., Ladopoulos, V., Mitchelmore, J., Goode, D. K., Calero-Nieto, F. J., Moignard, V., et al. (2016). Integrated genome-scale analysis of the transcriptional regulatory landscape in a blood stem/progenitor cell model. *Blood*, 127(13):e12–e23.

Wilson, N. K., Timms, R. T., Kinston, S. J., Cheng, Y.-H., Oram, S. H., Landry, J.-R., Mullender, J., Ottersbach, K., and Gottgens, B. (2010b). Gfi1 expression is controlled by five distinct regulatory regions spread over 100 kilobases, with scl/tal1, gata2, pu. 1, erg, meis1, and runx1 acting as upstream regulators in early hematopoietic cells. *Molecular and cellular biology*, 30(15):3853–3863.

Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, 24(1):238–241.

Wutz, G., Varnai, C., Nagasaka, K., Cisneros, D. A., Stocsits, R. R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M. J., Walther, N., Koch, B., Kueblbeck, M., Ellenberg, J., Zuber, J., Fraser, P., and Peters, J. M. (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.*, 36(24):3573–3599.

Xiong, N., K. C. and Raulet, D. H. (2002). Redundant and unique roles of two enhancer elements in the tcrgamma locus in gene regulation and gammadelta t cell development. *Immunity*, 16(3):453–63.

Xu, C. and Corces, V. G. (2018). Nascent DNA methylome mapping reveals inheritance of hemimethylation at CTCF/cohesin sites. *Science*, 359(6380):1166–1170.

Yao, Y., Minor, P. J., Zhao, Y. T., Jeong, Y., Pani, A. M., King, A. N., Symmons, O., Gan, L., Cardoso, W. V., Spitz, F., Lowe, C. J., and Epstein, D. J. (2016). Cis-regulatory architecture of a brain signaling center predates the origin of chordates. *Nat. Genet.*, 48(5):575–580.

Yokomizo, T., Hasegawa, K., Ishitobi, H., Osato, M., Ema, M., Ito, Y., Yamamoto, M., and Takahashi, S. (2008). Runx1 is involved in primitive erythropoiesis in the mouse. *Blood*, 111(8):4075–4080.

Yousefzadeh, M. J., Wyatt, D. W., Takata, K., Mu, Y., Hensley, S. C., Tomida, J., Bylund, G. O., Doublie, S., Johansson, E., Ramsden, D. A., McBride, K. M., and Wood, R. D. (2014). Mechanism of suppression of chromosomal instability by DNA polymerase POLQ. *PLoS Genet.*, 10(10):e1004654.

Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., Tian, S., Nie, J., Jonsdottir, G. A., Ruotti, V., Stewart, R., Slukvin, I. I., and Thomson, J. A. (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318(5858):1917–1920.

Yu, M., Mazor, T., Huang, H., Huang, H. T., Kathrein, K. L., Woo, A. J., Chouinard, C. R., Labadorf, A., Akie, T. E., Moran, T. B., Xie, H., Zacharek, S., Taniuchi, I., Roeder, R. G., Kim, C. F., Zon, L. I., Fraenkel, E., and Cantor, A. B. (2012). Direct recruitment of polycomb repressive complex 1 to chromatin by core binding transcription factors. *Mol. Cell*, 45(3):330–343.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., et al. (2014). A comparative encyclopedia of dna elements in the mouse genome. *Nature*, 515(7527):355–364.

Yusufzai, T. M., Tagami, H., Nakatani, Y., and Felsenfeld, G. (2004). CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol. Cell*, 13(2):291–298.

Yzaguirre, A. D., de Bruijn, M. F., and Speck, N. A. (2017). The role of runx1 in embryonic blood cell formation. In *RUNX proteins in development and cancer*, pages 47–64. Springer.

Zhang, P., He, Q., Chen, D., Liu, W., Wang, L., Zhang, C., Ma, D., Li, W., Liu, B., and Liu, F. (2015). G protein-coupled receptor 183 facilitates endothelial-to-hematopoietic transition via notch1 inhibition. *Cell research*, 25(10):1093–1107.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., et al. (2008). Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):1.

Zhao, H. and Dean, A. (2004). An insulator blocks spreading of histone acetylation and interferes with RNA polymerase II transfer between an enhancer and gene. *Nucleic Acids Res.*, 32(16):4903–4919.

Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., and Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.*, 38(11):1341–1347.

Zheng, H. and Xie, W. (2019). The role of 3D genome organization in development and cell differentiation. *Nat. Rev. Mol. Cell Biol.*, 20(9):535–550.

Zhou, J., Shen, B., Zhang, W., Wang, J., Yang, J., Chen, L., Zhang, N., Zhu, K., Xu, J., Hu, B., Leng, Q., and Huang, X. (2014). One-step generation of different immunodeficient mice with multiple gene modifications by CRISPR/Cas9 mediated genome engineering. *Int. J. Biochem. Cell Biol.*, 46:49–55.

Zhou, W., Dinh, H. Q., Ramjan, Z., Weisenberger, D. J., Nicolet, C. M., Shen, H., Laird, P. W., and Berman, B. P. (2018). DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.*, 50(4):591–602.

Zhu, J. and Paul, W. E. (2008). Cd4 t cells: fates, functions, and faults. *Blood*, 112(5):1557–69.

Zuniga, A. and Zeller, R. (2014). Development. in turing's hands–the making of digits. *Science*, 345(6196):516–7.