The functional and spatial organization of chromatin during Thymocyte development

Yousra Ben Zouari

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

IGBMC - CNRS UMR 7104 - Inserm U 1258

THÉSE présentée par:

Yousra BEN ZOUARI

Soutenue le : **03 May 2018**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Aspects moléculaires et cellulaires de la biologie

The functional and spatial organization of chromatin during Thymocyte development

THÈSE dirigée par :

M. SEXTON Thomas	CR, IGBMC, Illkirch, France

RAPPORTEURS :

Mme SACCANI Simona	DR, IRCAN, Nice, France
M. MANKE Thomas	Professeur, Max Planck, Freiburg, Allemagne

AUTRES MEMBRES DU JURY :

Mme BOEVA Valentina	CR, Institut Cochin, France
M. KASTNER Philipe	MCF, IGBMC, Illkirch, France
Mme SOUTOGLOU Evi	DR, IGBMC, Illkirch, France

Acknowledgements

First of all, I would first like to thank my thesis advisor Dr. Siomona Saccani, Dr. Philippe Kastner, Professeur Thomas Manke, Dr. Valentina Boeva and Dr. Evi Soutoglou for accepting to be members of my thesis committee. I am sure your expertise will provide great insight and help defining future directions of the project.

Afterwards, I want to a give special thanks to my supervisor Tom Sexton who made these almost 4 years of PhD an exciting experience. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly. Thank you for being supportive, trustful and encouraging. I am very grateful for everything you did for me.

I would like to thank both past and current members of Sexton laboratory. Anne, that started the journey with me, thank you for the laughs and fun moments and for your support in the hard times. I need more than these lines to tell you how much I'm counting on you my best friend. Dominique, you are the wisdom and the caring of the lab. Thank you for being helpful and supportive. I am grateful to every advice you gave me. Audrey, thank you for the joy and fun we had together. I was very pleased to share with you the same space. Sunjay, thank you for the Indian food you made for every special moments. Nezih, thank you for driving me all the time and for the fun moments we had in the car with Turkish music without forgetting the good food of your mum. Manon, thank you for your pure smile and your optimism in the hard times. Natalia, thank you for your smile and spoiling us with chocolates. Angeleki, thank you for being so pretty and elegant. You are the sunshine of our lab😊.

I would like also to thank other people from IGBMC who became very close friends during these past 4 years. Rana and Naima, thank you for your kindness and the joyful moments I had with you. It was a great pleasure for me to spend time with you.

I also have to thank everyone that, one way or another, had an important technical and/or scientific input on my work. I want to thank everyone from the IGBMC Informatic servicesn namely Serge Uge for the help with server problems.

Je souhaite également remercier ma famille pour leur soutien et leur amour. Maman et papa, Fathia et Mokhtar, je ne trouve pas de mots pour vous dire à quel point je vous suis reconnaissante pour tout ce que vous avez et continuez à faire pour moi. San nul doute, c'est grâce à vous que je suis la personne que je suis aujourd'hui. Merci pour votre amour à toute épreuve et pour avoir toujours fait en sorte que je ne manque de rien. Mes sœurs et mon frère, Intissar, safa et yassine, je ne pouvais pas rêver de meilleur frère et sœur. J'ai beaucoup de chance d'avoir partagé mon enfance avec vous.

Je remercie Sameh, mon meilleur ami pour avoir toujours été là. Aussi loin que je m'en souvienne, tu as toujours su trouver les mots pour me remonter le moral pour calmer mes crises de panique. Il me faudrait bien plus que ces quelques lignes pour te dire à quel point je compte sur toi.

Je remercie également mes amis Afef et Takoi pour toutes les soirées, journées, sorties etc … Vous êtes des amis comme tout le monde en rêverait ! Je sais que je peux compter sur vous aussi bien pour faire la fête qu'en cas de besoin !

Finally, I would like to thank also Badr, for being supportive, encouraging and caring. You have been a part of this dream which came true. Thank you for being a part of this journey and pushing me forward in the hard times.

SUMMARY

Introduction

DNA, the genetic code of nearly all living organisms, is associated with proteins, predominantly histones in eukaryotes to facilitate their folding into chromatin within the cell nucleus. Chromatin needs to be densely compacted, while still allowing access of genes and regulatory elements for control of biological processes such as transcription. Chromosome folding takes place at different hierarchical levels, with various topologies correlated with control of gene expression. At the kilobase-to-megabase scale, chromatin forms loops which bring gene promoters and their distal regulatory elements, such as transcriptionally activating enhancers, into direct physical proximity. These topologies have been proposed to form an "active chromatin hub", whereby the combination of regulatory factors bound to the promoter and interacting enhancers generates an environment permissive to transcriptional activation. At a higher level, groups of chromatin loops are confined within larger, megabase-scale structures termed "topologically associated domains" (TADs). These can in turn be organized into transcriptionally active ("A") or repressed ("B") compartments. TAD borders may play an important functional role in preventing aberrant contacts between enhancers and non-target neighboring genes.

Despite the large number of recent studies describing chromatin topologies and their correlations with gene activity, many questions remain, in particular how these topologies are formed and maintained. Interestingly, the epigenetic state (as determined by histone modifications) of enhancers varies much more widely across cell types than at promoters, suggesting that most cell fate potential is actually encoded at a distance from developmental genes. Studies of chromatin loops between developmental genes and their enhancers give different views on the role of chromatin topology in gene control. Some indicate that

chromatin loops form concomitantly with transcriptional activation, concluding that the topology is directly responsible for gene expression. However other studies show that the chromatin loop can precede transcriptional activation, suggesting that the topology represents an earlier "priming" event, allowing the gene to be regulated by downstream acute signals. Furthermore, it is not clear how and which epigenetic marks on enhancers may be instructive for chromatin loop formation. Analogously, questions remain as to what extent TAD structures are developmentally plastic or stable. It is thus very important to understand better the link between epigenetic marks, chromatin topology and transcriptional control.

Most studies of chromatin topology are based on the chromosome conformation capture (3C) method, whereby formaldehyde crosslinked chromatin regions is digested and re-ligated to create chimeric DNA products between restriction fragments which were physically proximal *in vivo*. Specific target products can be amplified and measured by quantitative PCR; the Hi-C method employs high-throughput sequencing to measure such chromatin interactions on a genome-wide scale. However, the strength of this latter approach is also its disadvantage: the number of potential interactions which can be detected is far greater than current sequencing capacities, limiting the resolution of the technique. To overcome this limitation, we and others have introduced an oligonucleotide capture step into the Hi-C procedure (Cap-C) to study interactions within subsets of genomic regions at high resolution. In one experimental setup, we have designed capture probes to each of the ~22,000 mouse promoters to systematically characterize their chromatin looping interactions. In another, we have designed a tiling strategy across selected ~600 kb regions, targeting studies of chromatin interactions to those flanking TAD borders close to differentially expressed developmental genes.

In the lab, we performed Cap-C experiments to study topological changes during thymocyte development, specifically at the transition between Double Negative (DN) and

Double Positive (DP) cells, representing the critical checkpoint for productive T cell receptor gene rearrangement. By linking chromatin topology dynamics with known transcriptomic and epigenomic changes during this transition, we aim to address the following questions:

1. Is chromatin structure stable or dynamic during development?
2. How are chromatin topologies (loops and TADs) established and maintained?

Methods

During my PhD, I performed the computational analyses of the lab's Cap-C datasets. Although these datasets have a superior resolution to most Hi-C datasets, the previously developed bioinformatics tools for Hi-C analysis were inappropriate for the unique challenges presented by Cap-C data. To determine the most reliable interactions between promoters and distal regulatory elements, I developed a computational method (PromoMaxima) more robust and specific than previously published approaches. I then performed integrative analyses of these called interactions (and called TADs/borders from the other Cap-C strategy) with published ChIP-seq and RNA-seq data, to obtain the following hypotheses and conclusions about the interplay of genome structure and function in thymocyte development.

Results

I. Identification of a complex network of dynamic chromatin loops during thymopoesis

With PromoMaxima, I identified thousands of chromatin looping interactions in thymocytes. Whereas many were developmentally stable, hundreds were nevertheless much stronger in one cell type or another, often linked to a transcriptional change of the interacting gene. As was previously reported, a large number of chromatin loops were detected between promoters and distal regions containing CTCF binding sites and/or the histone signatures of enhancers. Although many epigenomic studies have distinguished the histone modifications

of “poised” and “active” enhancer elements, such marks do not seem to be predictive of looping state: depending on the genomic context, poised or active enhancers appear just as likely to interact with their target genes. However, unlike the previous studies, which focused on enhancers, we also identified hundreds of interactions with distal elements which correlated with *repression* of the target gene, indicating an extensive network of distal silencers. These regulatory elements have been described in specific case studies decades ago, but to date no study has characterized them on a genome-wide scale, nor identified a signature epigenomic mark. The putative silencers I identified are also enriched for CTCF sites, but are depleted of active histone modifications, and enriched in LINE repetitive elements. Other members of the lab have already functionally validated a number of the putative silencers that I identified, and are currently performing experiments to characterize them in greater depth *in vivo*.

II. TADs are predominantly developmentally stable, with notable remodeling at specific borders.

In contrast with our findings for chromatin loops, TADs appear very robust to developmental changes, with the structures largely maintained despite large transcriptional changes in the underlying genes. However a minority of TADs were remodeled during the DN-to-DP transition, in each case linked to transcriptional induction of the component genes. We observed: 1) The formation of new “sub-TADs” containing the body of the induced gene; 2) a shift in the border of an existing TAD, so that the enhancer is in the same TAD as the entire transcribed gene, and not just the poised promoter. Artificial transcriptional induction of these genes by the dCas9-VP64 system showed that transcription was sufficient to remodel TADs in some cases but not others.

Conclusion and discussion

The recently developed Cap-C technique, optimized within the lab and coupled with the new analytical method I developed, allows efficient and sensitive detection of looping chromatin interactions. We have uncovered extensive chromatin topology dynamics during thymocyte development, much of which is correlated with transcriptional regulation. In particular, we uncovered networks of interactions with putative regulatory elements, both activating enhancers *and* repressing silencers, the latter at a previously underappreciated scale. Previous studies have noted an enrichment of SINE repetitive elements at enhancers, and have hypothesized that these and long terminal repeat retroviral activating elements could have been co-opted during evolution to activate endogenous genes. Based on our finding of LINE enrichment at putative silencers, it is interesting to speculate that these regions, normally shut down by host defense mechanisms against ancestral parasitic elements, may also be co-opted as developmental repressive elements. Future experiments in the lab will explore this possibility, and their potential interplay with the CTCF sites with which they are juxtaposed.

Very recent studies have given conflicting information on whether transcription directly instructs TAD formation or remodeling. We have shown that the majority of TADs are robust to transcriptional changes during development, but that a subset are reorganized around induced genes, in some cases directly. Future experiments of the lab will examine mechanisms other than transcription which may influence chromatin architecture, such as differential binding of CTCF, and how these may interplay with transcriptional control and chromatin architecture.

Résumé

Introduction

L'ADN constitue le patrimoine génétique de la plupart des organismes vivants. Il est associé à des protéines dont majoritairement des histones pour former la composante principale du noyau, la chromatine. Celle-ci est fortement condensée pour tenir dans le noyau, une organisation génomique complexe qui toutefois permet l'accessibilité de l'ADN aux différentes activités nucléaires. Ainsi, le contrôle de la transcription survient dans un contexte de repliement chromosomique avec différents niveaux hiérarchiques. A l'échelle de plusieurs centaines de kilobases, la chromatine forme des boucles qui permettent les contacts physiques à distance entre les amplificateurs de transcription « enhancers » et les promoteurs de leurs gènes cibles. Ces structures de chromatine forment ainsi des « active chromatin hubs » qui amènent les facteurs de transcription à se lier aux promoteurs et aux éléments enhancers formant un environnement de régulation plus permissif que celui des promoteurs isolés. Le second niveau hiérarchique est constitué d'un ensemble de boucles chromatiniennes confinées dans des structures de l'ordre du mégabase appelées « domaines topologiques ». Selon l'activité des gènes inclus, les domaines topologiques constituent ensemble un de compartiments actifs « A » ou inactifs « B ». Les frontières de ces domaines topologiques jouent le rôle de barrière en empêchant les contacts aberrants entre des éléments régulateurs et des gènes voisins.

Malgré les vastes études démontrant le rôle de la conformation génomique dans le contrôle transcriptionnel, de nombreuses questions restent en suspens, et en particulier, comment ces structures chromatiniennes sont formées et maintenues. De manière intéressante, l'état de la chromatine au niveau des séquences enhancers varie bien plus d'un type cellulaire à l'autre que celui des promoteurs de gènes, suggérant que le potentiel de régulation

épigénétique est principalement porté par les enhancers. La plupart des modèles qui cherchent à expliquer le rôle des enhancers impliquent des boucles de chromatine, rapprochant les séquences enhancers avec les régions promotrices des gènes. Certains indiquent que la boucle de chromatine se forme de manière concomitante à l'activation de la transcription et concluent que les interactions enhancer-promoteur stimulent directement l'expression des gènes. D'autres montrent que la boucle de chromatine en réalité précède la transcription suggérant que les structures formées sont des événements épigénétiques déjà présents rendant le locus compétent pour une expression efficace en réponse à des signaux de développement tardif. De plus, il n'est pas clair si les profils épigénétiques différents au niveau des enhancers affectent la capacité de former des interactions avec les gènes cibles. Des études précédentes proposent également des points de vue conflictuels à propos de la maintenance des domaines topologiques durant la différenciation cellulaire. Certains montrent que les domaines topologiques sont des structures stables en se basant sur une étude exhaustive de la conformation génomique de différent type cellulaire. D'autres les décrivent comme des structures dynamiques. Il est donc primordial de mieux comprendre les liens entre **l'état de la chromatine au niveau des éléments régulateurs, la topologie de la chromatine et la régulation de la transcription.**

L'étude de l'organisation spatiale des chromosomes est basée sur une approche de capture de la conformation chromosomique (3C). Cette technique permet de lier entre elles, grâce au formaldéhyde, les zones chromosomiques proches. Les étapes de digestion / ligation permettront finalement de révéler les rapprochements qui seront détectés par PCR quantitative. Quant au séquençage à haut débit, il donnera accès aux repliements chromosomiques à l'échelle du génome, on parle alors de Hi-C. Cependant, la force de cette approche dans l'accessibilité à toutes les interactions possibles est également sa faiblesse : le nombre d'interactions qui devrait être détecté est bien supérieur à la capacité actuelle de

séquençage, conduisant à une perte d'information. Pour contourner ces limitations, nous avons introduit une étape supplémentaire de capture de séquences dans la procédure du Hi-C pour augmenter la résolution à un sous-ensemble de régions du génome (Cap-C). Ainsi, nous avons utilisé des sondes de capture pour les 22 000 promoteurs des gènes de la souris, afin de caractériser systématiquement les interactions chromosomiques entre tous les promoteurs de gènes et leurs enhancers. Dans une deuxième série d'expérience, nous avons ciblés quelques frontières des domaines topologiques contenant des variations d'expression génique essentielles au cours de processus de développement.

Ces expériences de capture de la conformation chromosomique ont été réalisées pour le processus de différenciation des thymocytes en tenant compte uniquement des stades développementaux critiques : Double Négatif (DN) et Double positif (DP). Nous espérons mettre en évidence les liens entre la conformation de la chromatine avec le contrôle de l'expression génique tout en répondant aux questions suivantes :

1.1 La structure chromatinienne, est-elle stable ou dynamique durant la différenciation cellulaire ?

1.2 Comment les structures chromatiniennes (domaines topologiques et boucles chromatiniennes) sont-elles formées et maintenues ?

Méthodologie

Durant ma thèse, j'ai été en charge de l'analyse des données issues des expériences de Cap-C. Les Cap-Cs ont montré une résolution bien supérieure à celles des HiC, cependant les outils d'analyse bio-informatique disponibles se sont avérés inappropriés. Afin de déterminer les interactions significatives entre les promoteurs et les éléments régulateurs, j'ai donc développé une méthode d'analyse plus robuste et efficace que les approches déjà publiées. Par ailleurs, j'ai analysé et intégré les données de Chip-Seq et RNA-seq avec les données de

structure chromatinienne afin de comprendre le lien entre la conformation des chromosomes et la régulation des gènes tant sur le plan épigénétique que transcriptionel.

Résultat

I. Identification d'un large éventail de boucles chromatiniennes au cours du développement des thymocytes

Grâce à notre nouvelle approche, j'ai identifié des milliers de boucles chromatiniennes. Nous avons pu observer que la majorité de ces boucles sont stables au cours du processus de développement des thymocytes. Un certain nombre d'entre elles présente néanmoins un profil dynamique, souvent liées avec une réponse transcriptionelle du gène cible. Comme il a également été publié, un grand nombre de ces boucles ont été répertoriées entre les promoteurs et les régions régulatrices qui portent la signature chromatiniennes des enhancers ainsi que des sites de liaison de CTCF. Bien que de nombreuses études en épigénomique ont identifiées des marques distinctes d'histones entre les enhancers actifs et « poised » pour l'activation des gènes à différentes étapes du développement, ces marques épigénétiques ne sont pas prédictives de la formation des boucles de chromatine. En effet, selon le contexte génomique, tant les enhancers actifs que les enhancers « poised » participent à la formation des boucles chromatiniennes en liant les promoteurs cibles. Contrairement aux études antérieurs qui se sont focalisées sur les enhancers, nous avons pu déterminer des nouveaux éléments régulateurs impliqués dans la *répression* de l'expression (les « silencers »). Cette classe d'élément régulateur a été décrite il y a quelques décennies, mais aucune étude n'a pu les caractériser à l'échelle du génome jusqu'à présent. Le profil épigénétique des silencers se distingue par une absence de marqueurs d'histone active et es enrichi par la présence d'éléments répétitifs de la classe des LINEs. L'équipe a déjà réalisé un certain nombre de

validation de nouveaux silencers identifié par mes soins, et a présente tente de les caractériser plus en profondeur *in vivo*.

II. Les domaines topologiques sont des structures stables avec quelques changements potentiels au niveau de leurs frontières

En revanche, les domaines topologiques semblent être des structures robustes sur le plan développemental, avec très peu de changements observés entre les deux types cellulaires. Une minorité de domaines ont été remodelés au cours du développement, liés à l'induction transcriptionelle des gènes. Nous avons observé : 1) la formation des nouveaux domaines sur des gènes transcrit. 2) un shift de frontière d'un domaine topologique afin d'inclure la boucle en chromatine du promoteur et son enhancer. L'induction artificielle de ces gènes a montré que certains changements de TAD peuvent être liés à la transcription, tandis que d'autres ne le sont pas.

Conclusion et Discussion

La technique des Cap -C est récente ce qui explique que le laboratoire a dû mettre au point des outils d'analyses complémentaires afin de déterminer de manière fiable les interactions chromatinienne au niveau des régions ciblées. La méthode d'analyse, que j'ai établie, a démontré une bonne efficacité et sensibilité pour la détection de ces interactions chromatiniennes et permettra donc de répondre de manière plus précise aux questions biologiques posées. Nous avons ainsi pu décrypter la structure chromatinienne associée à la différenciation des thymocytes et mettre en évidence des mécanismes de contrôle transcriptionnel de certains gènes. Nous avons identifiés différents éléments régulateurs dont les enhancers et les silencers. Par ailleurs, des études déjà publiées ont montré une corrélation de la présence d'éléments de répétitions SINEs à proximité des enhancers. Dans notre

approche, nous avons pu vérifier ses observations et nous avons également mis en évidence une corrélation entre les éléments LINEs et les silencers, d'autre part. Il est intéressant de s'interroger sur les éléments répétitifs du génome. En effet, ils sont considérés comme des « éléments parasitaires ancestraux » qui peuvent être utilisés au cours de l'évolution pour le contrôle des gènes développementaux. Ainsi, il a été proposé que des enhancers rétroviraux ancestraux participent à l'activation de gènes et que d'autres classes d'éléments répétitifs, qui sont naturellement réduites au silence dans le cadre de la défense du génome hôte contre la transposition, puissent aussi être co-optées pour la répression des gènes.

Des études très récentes ont montré des conclusions contradictoires sur la question de savoir si la transcription est directement liée à la formation de domaines topologiques. Nous avons démontré que la plupart des domaines sont robustes aux changements de la transcription, mais qu'il y a certains domaines topologiques qui peuvent être réorganisés directement suite à l'induction des gènes. Les expériences futures de l'équipe vont consister à examiner les facteurs (hors transcription) qui peuvent influencer l'architecture de la chromatine, comme la liaison différentielle des CTCF, et comment ces facteurs peuvent être coordonnés par le contrôle de transcription.

Table of contents

List of figures

INTRODUCTION

RESULTS

I. Hi-C and CHi-C quality control

II. PromoMaxima: a pipeline for detection and visualization of *cis*-DNA looping in Capture Hi-C

III. Developmentally dynamic gene promoter interactions in transcriptional activation and repression

Figure 5: The mouse thymocyte and ES promoter interactome features.

Figure 6: Stable thymocyte promoter interactions.

Figure 7: Dynamic thymocyte promoter interactions.

Figure 8: Stable and dynamic promoter interactions linked to transcriptional activation and repression.

Figure 9: Following up the link between LINEs and putative silencers.

IV. TADs caller benchmarking

Figure 1: Comparative results of methods for the identification of TADs.

V. Transcription directly remodels a small subset of topologically associated domains

Figure 1: Conservation of TAD structure across thymocyte development.

Figure 2: Transcription-coupled sub-TAD formation at the *Nfatc3* gene.

Figure 3: Transcription-coupled sub-TAD formation at the *Tmem131* gene.

Figure 4: Transcription directly remodels the *Nfatc3* sub-TAD.

Figure 5: TAD border remodeling around the *Bcl6* gene during thymocyte maturation.

Figure 6: CHi-C enhances resolution of interaction maps.

Figure 7: High reproducibility across biological replicates.

Figure 8: Potential differential CTCF binding at the *Bcl6* locus.

Figure 9: Transcriptional induction does not remodel the *Bcl6* TAD in ES cells.

Figure 10: TAD remodeling events uncovered in Hi-C.

List of tables

RESULTS

I. Hi-C and CHi-C quality control

II. PromoMaxima: a pipeline for detection and visualization of *cis*-DNA looping in Capture Hi-C

III. Developmentally dynamic gene promoter interactions in transcriptional activation and repression

V. Transcription directly remodels a small subset of topologically associated domains

INTRODUCTION

Introduction

The DNA is the genetic material that encodes for all the information essential for life (Dahm, 2005). In eukaryotes, it is wrapped around structural proteins called histones to construct strings of nucleosomes that can be further compacted into a three-dimensional (3D) organization within cell nuclei. At the most extreme, during metaphase, the chromatin fiber folds to the 0.7 µm thick chromatid. The process underlying this compaction remains unclear, although condensins and topoisomerase IIα are implicated in this process (Swedlow and Hirano, 2003;Gibcus *et al.*, 2018). Even at interphase, the spatial arrangement of the chromatin in the nucleus is highly organized at different levels, and can have a direct impact on genomic activity, such as transcription, by regulating DNA accessibility to the genomic machinery. For example, histones can impede the access of many regulatory proteins to their binding motifs, and hinder the movement of polymerases along the DNA fiber. The post-translational modification of histone tails, and/or chromatin remodeling on binding of sequence-specific transcription factors, can facilitate access to DNA, in turn activating some genomic elements (Berger, 2007). Different studies recently demonstrated a further correlation between chromatin topology and underlying gene activity (Cavalli & Misteli, 2013). For example, it was revealed that chromatin looping events can facilitate transcription by bringing distal regulatory elements such as enhancers in direct physical proximity with gene promoters (Palstra et al., 2003). Developmental fate decisions are underpinned by the combinatorial action of tissue-specific enhancers (Osterwalder et al., 2018); it is therefore likely that promoter-enhancer interactions need to be highly regulated to prevent aberrant gene responses. At the megabase scale, the genome folds into discrete 3D structures that tend to favor internal rather than external interactions. These structures have been termed

"topologically associating domains" (TADs) and they are largely conserved among different cell types in animals (Sexton & Cavalli, 2015). At the chromosome level, each chromosome occupies different nuclear regions termed chromosome territories, which are radially organized such that gene-poor chromosomes are placed at the nuclear periphery and the gene-rich chromosomes occupy more central positions (Cremer & Cremer, 2001). Over the past decades many different technologies have been developed in order to assess genome organization, the principles underlying its folding and its relationship with its activity. However it is still unclear whether chromosome folding is a cause or a consequence of genomic functions. In this Introduction, I will describe our current understanding of the link between gene position or chromosome folding and the potential for transcriptional regulation, before giving a technical appraisal of the different methods that have allowed us to interrogate chromosome folding. As our group uses thymocyte differentiation as a model system for studying developmental dynamics of chromatin topology, I will then give a description of this process, and then highlight the Research Aims of my thesis in the following section.

I: Nuclear and genome architecture

1. An overview of nuclear organization

Since early microscopy studies identified the partitioning of chromatin into densely-packed heterochromatin and lighter-staining euchromatin, it has been appreciated that the nucleus is a highly heterogeneous organelle, likely linked to regulation of the underlying genes. In this section, I discuss nuclear substructures which have been implicated in transcriptional regulation.

1.1 The nuclear periphery

With rare exceptions (Solovei et al., 2009), heterochromatin is predominantly located at the periphery of the nucleus, which is proposed to form a repressive environment due to restricted access of transcription factors and polymerase to DNA sequences. In support of this, gene-poor chromosomes preferentially occupy more peripheral radial locations in the nucleus (Cremer & Cremer, 2001), and specific genes can relocate from the periphery to the nuclear interior on transcriptional induction (Chuang et al., 2006; Kosak et al., 2002). One factor implicated in gene repression at the periphery is the nuclear lamina, an architectural support for the internal nuclear membrane. It is composed of intermediate filament proteins (nuclear lamins). The lamins interact with different repressive chromatin proteins, in particular heterochromatin protein 1 (HP1) (Ye, Callebaut, Pezhman, Courvalin, & Worman, 1997) and histone deacetylases (Somech et al., 2005). Genome-wide approaches have identified large genomic regions (lamin-associated domains; LADs) which associate with the lamina (Peric-Hupkes et al., 2010). In general, LADs are associated with repressed transcription, which may be directly caused by lamin interactions and/or attachment of the chromatin to the nuclear periphery. We distinguish two types of LADs: cell type specific LADs and conserved LADs (Meuleman et al., 2013). The conserved LADs usually span gene poor genomic regions with low GC content, whereas cell type-specific LADs span genomic regions that enclose tissue-specific genes. It is currently unclear if such facultative lamina attachment is a direct cause of transcriptional repression of these developmental genes. For example, the artificial tethering

of genes to the lamina did not always result in transcriptional silencing (Finlan et al., 2008; Reddy, Zullo, Bertolino, & Singh, 2008).

1.2 The nuclear pore complex

Not all regions of the nuclear periphery are necessarily repressive. The nuclear pore complex is an evolutionarily conserved structure regulating all transport of protein and mRNA between the nucleus and the cytoplasm, but it also appears to play a role in cell division and transcriptional activation (Ptak, Aitchison, & Wozniak, 2014). Electron microscopy studies in yeast demonstrated the presence of transcriptionally active regions (euchromatin) around the nuclear pore complex while the heterochromatin regions were adjacent to the nuclear lamina (Rodríguez-Navarro et al., 2004). This suggests that the nuclear pore complex may be involved in the activation of transcription, and/or facilitates efficient export of nascent mRNA to the cytosol for translation (Capelson et al., 2010). Most evidence for the role of the nuclear pore complex in transcriptional control has been obtained in yeast; it is unclear whether similar mechanisms are conserved in species with much larger nuclei, where chromatin access to the periphery may be more limited. In *Drosophila*, nuclear pore components (nucleoporins) have been implicated in dosage compensation (Mendjan et al., 2006), and mammalian nucleoporins have been shown to be involved in diverse activities, including gene activation (Ptak et al., 2014), but it is unclear whether such activities occur at genuine nuclear pores or different nucleoplasmic protein complexes containing nucleoporins.

1.3 The nucleolus and other nuclear foci

The nucleolus is a ribosome production "factory" where the rRNA is transcribed and the ribosomal subunits are assembled. It is usually organized around the genomic regions that contain rRNA genes and transcribed by RNA polymerase I (PolI) (Németh et al., 2010). Curiously, this highly active nuclear landmark is frequently surrounded by perinucleolar heterochromatin. Recent studies have identified DNA sequences bound to biochemically isolated nucleoli (nucleolus-associated domains; NADs) (van Koningsbruggen et al., 2010). They comprise large domains interspersed across all the chromosomes, including those lacking rDNA loci (van Koningsbruggen et al., 2010). Generally, NADs are AT-rich and gene-poor, covering about 4% of the human genome which includes tissue-specific repressed regions, transposable elements and repetitive sequences (Thomson, Gilchrist, Bickmore, & Chubb, 2004). Some genes found to associate with the periphery of the nucleus (namely, LADs) were shown also to associate at the nucleolus, such as olfactory receptor genes (Clowney et al., 2012). Since nucleoli are not found at the periphery, this implies a heterogeneous nuclear organization within cell populations, whereby many loci can be repressed equally well at either the perinucleolar environment or the lamina.

In addition to rRNA, mRNA transcription also appears to be highly compartmentalized in the nucleus. Labeling of RNA polymerase II or nascent RNA revealed that virtually all gene transcription takes place in a relatively limited number of foci or "transcription factories" (Jackson & Cook, 1985; Osborne et al., 2004). Active genes have

been shown to colocalize at factories, presumably for their efficient co-regulation. In support of this, there is evidence that genes sharing common transcription factors may preferentially co-occupy "specialized factories" enriched in these factors(Papantonis et al., 2012; Schoenfelder et al., 2010). However, recent super-resolution microscopy and live imaging experiments raise questions as to how ubiquitous and/or stable such factories may be (Cisse et al., 2013; Conic et al., 2018; Zhao et al., 2006).

In *Drosophila*, co-regulated gene clustering has additionally been described for repressed genes, which are recruited to foci of Polycomb group protein repressors (Bantignies et al., 2011), implying the existence of silent spatial gene networks as well as active ones. Although the existence of such "Polycomb bodies" is contentious in mammals (Saurin et al., 1998), a growing body of evidence in mouse embryonic stem (ES) cells suggests that many Polycomb-regulated genes spatially co-associate in networks distinct from those linked to pluripotency transcription factors (Denholtz et al., 2013; Schoenfelder, Sugar, et al., 2015).

2. Chromosome organization in the nuclear space

The nucleus carries many structural features, some of which have been observed in microscopy studies since the early twentieth century. However, with the advent of the chromosome conformation capture (3C) technology and its derivatives (see section II - 2.3 for details), the structural organization of the genome itself is now beginning to be appreciated.

Chromosomes appear to be hierarchically built up, with architectural features at each scale correlated with transcriptional control (Sexton & Cavalli, 2015).

2.1 Chromatin loops and gene regulation

2.1.1 *Cis*-regulatory elements: enhancers, silencers and insulators

Since early transgenic studies, it is appreciated that promoters alone are incapable of fully and sufficiently activating genes, particularly those implicated in cell development (Talbot et al., 1989) . For efficient gene transcription, some regulatory DNA regions that are distant from promoters are implicated, the best studied class of which is enhancers, which stimulate transcription. Most metazoan genes are under the control of these enhancers, which can act over megabase distances, and even from within introns of unregulated genes (Amano et al., 2009). The first enhancer identified was a 72 bp element of the SV40 virus genome which was capable of activating the transcription of a reporter gene in HeLa cells (cancer cell line) by several hundred-fold (E. May, Omilli, Ernoult-Lange, Zenke, & Chambon, 1987) . Since then, transgenic experiments and reporter assays genetically identified many enhancers as short DNA motifs that act as binding sites for specific transcription factors, which activate transcription independently of the distance and orientation of their target gene. Recently, elegant genome-wide versions of such reporter assays, such as STARR-seq (Arnold et al., 2013), allow identification of enhancer elements within specific mammalian cell types (Muerdter et al., 2017; Vanhille et al., 2015). A large body of epigenomic profiling studies have correlated enhancers with signature chromatin features, such as histone lysine 4 monomethylation (H3K4me1), H3K27 acetylation (H3K27ac), hypersensitivity to DNaseI digestion, and the production of short bidirectional transcripts (eRNAs) (Creyghton et al., 2010; Heintzman et al., 2009; Kim et al., 2010; Koch et al., 2011; Rada-Iglesias et al., 2011).

Enhancers lacking these extra features, and sometimes even encompassing repressive marks, such as H3K27 trimethylation (H3K27me3), are proposed to be "poised" enhancers, which may become activated at later developmental stage, or "decommissioned" enhancers, which were active in prior stages.

Silencers are the functional opposite of enhancers, defined as genetic elements which negatively regulate gene transcription in a position-independent fashion, and were first described more than three decades ago (Kadesch, Zervos, & Ruezinsky, 1986). Since then, several silencers have been discovered to control the expression of key developmental and immunological model genes (e.g. (Sawada, Scarborough, Killeen, & Littman, 1994)). However unlike for enhancers, no genome-wide identification of silencers has been made to date, and it is currently unknown how extensive they are, nor if they carry a signature epigenetic mark. Notably, very recent studies aimed at dissecting functional subsequences within selected enhancers have revealed that some can be bound by a spectrum of activating and repressing transcription factors, depending on the cellular context (Rajagopal et al., 2016). Thus, it is possible that some enhancers and silencers may comprise an overlapping set of genetic elements, which exhibit divergent behaviours under different biological conditions.

The third class of *cis*-regulatory element, insulators, does not directly activate or repress genes. Instead, they prevent communication between different genetic regions, defined by "enhancer-blocker" (preventing enhancer activation of a gene when placed in between them) or "barrier" (preventing spreading of heterochromatin) activities in genetic assays (West & Fraser, 2005). The predominant insulator in mammals is the binding motif for the factor CTCF (Phillips & Corces, 2009), although tRNA genes have also been described to have insulator activity (Raab et al., 2012).

2.1.2 Chromatin looping with enhancers

Until the advent of 3C, it was unclear how distal regulatory elements were able to exert their effects on target genes. Seminal studies of the beta-globin locus revealed that enhancers come into direct physical proximity with their target promoter by looping out the intervening chromatin (Palstra et al., 2003). The resulting "active chromatin hub" containing the regulatory factors at both the enhancer and the promoter is proposed to form a permissive environment for transcriptional firing. Numerous enhancer-promoter interactions have subsequently been identified in many different species and cell types. Notably, attempts to systematically identify all promoter-enhancer interactions within a given cell type (e.g. (Sanyal, Lajoie, Jain, & Dekker, 2012; Schoenfelder, Furlan-Magaril, et al., 2015a)) revealed that many enhancers do not contact (and presumably regulate) the genes that are closest on the linear chromosome fiber. However, it remains largely unknown exactly how enhancers find their cognate genes. One likely aspect dictating looping specificity is protein-protein interactions between compatible transcription factors bound to enhancer and promoter sequences. Initial studies in the beta-globin locus identified various erythrocyte-specific transcription factors, such as GATA-1, whose expression correlated with establishment of the enhancer-promoter loop (Drissen et al., 2004; Vakoc et al., 2005). Transcription factor exchange during development has also been associated with a rewiring of chromatin loops (Jing et al., 2008). Recent elegant experiments have even demonstrated that such protein-protein interactions can induce chromatin loops in certain contexts, which can even be causally linked to transcriptional induction (Deng et al., 2014; Morgan et al., 2017).

2.1.3 Chromatin looping NOT just transcription: the role of CTCF and cohesin

In addition to transcription factors, insulator proteins such as CTCF have been reported to be implicated in chromatin loop formation (Phillips & Corces, 2009), which are often stronger or

more stable than promoter-enhancer contacts (Rao et al., 2014a). Most of these CTCF mediated loops appear to be constitutive and associated with a more general architectural role, such as might be expected for a classical insulator preventing aberrant enhancer-promoter interactions (Phillips-Cremins et al., 2013a). However, the depletion of specific CTCF sites located right next to enhancers can actually perturb enhancer-promoter contacts and increase transcriptional noise. It thus appears that in these genomic contexts, CTCF-CTCF interactions are reinforcing enhancer-promoter interactions to confer robust expression control (Ren et al., 2017). Interestingly, the orientation of CTCF sites seems to be very important for loop formation. In fact, CTCF loops are almost exclusively between CTCF sites in convergent orientation (Rao et al., 2014a; Vietri Rudan et al., 2015). The disruption of CTCF orientation binding sites by inversion severely altered chromatin loops but it did not affect the CTCF binding (de Wit et al., 2015a; Guo et al., 2015). However, the inverted sites did not engage in *de novo* loops with compatible CTCF orientation, suggesting that other mechanisms dictate CTCF looping specificity.

Another major factor implicated in both transcriptional and architectural chromatin loops is cohesin, a multi-subunit protein complex initially recognized for its role in sister chromatid adherence, mitotic and meiotic chromosome segregation and DNA repair (Kim Nasmyth & Haering, 2009). Like CTCF, cohesin was also found to bind thousands sites of interphase chromatin but in a more tissue-specific manner (Parelho et al., 2008). In addition to that, cohesin was demonstrated to co-localize with the transcriptional co-activator Mediator and CTCF (Kagey et al., 2010), thus potentially facilitating enhancer-promoter and architectural looping. In fact, many of the original CTCF loops were later found associated to cohesin, and cohesin degradation severely disrupts all chromatin looping events (Rao et al., 2017). However, CTCF does not exclusively co-localize with cohesin and *vice versa*. Cohesin complexes have been shown to form rings to physically tether sister chromatids after DNA

replication (Kim Nasmyth & Haering, 2009); it is interesting to speculate that similar rings physically stabilize enhancer-promoter interactions, but this has yet to be demonstrated.

2.1.4 Chromatin loops - stable and/or dynamic structures?

In the beta-globin locus, only the expressed gene forms interactions with the enhancer, and only specifically in erythrocyte cells (Palstra et al., 2003), implying an instructive model (**Fig 1**) where chromatin looping is concomitant with, and necessary and sufficient for transcriptional induction. Subsequent studies made similar conclusions for many other genes (Sanyal et al., 2012; Schoenfelder, Furlan-Magaril, et al., 2015); for example, establishment of the promoter-enhancer loop at the endogenous *OCT4* locus distinguished reprogrammed from unresponsive cells during human induced pluripotent stem cell production (H. Zhang et al., 2013). However, other studies have identified pre-formed chromatin loops which can arise cell cycles before the target gene is transcribed, implying a permissive model where chromatin looping may be necessary but not sufficient for transcriptional firing (**Fig 1**). Examples of this instance include *Drosophila* mesoderm enhancers (Ghavi-Helm et al., 2014), and TNF-α responsive genes in human fibroblasts (Jin et al., 2013). This configuration has been proposed to allow rapid transcriptional induction of genes in response to acute stimuli, which is supported by the finding of paused RNA polymerase at many promoters participating in these "poised" interactions (Ghavi-Helm et al., 2014). The most recent systematic assessments of promoter-enhancer interactions during development actually found a high prevalence of both instructive and permissive loops (Freire-Pritchett et al., 2017; Rubin et al., 2017), but it is unclear what epigenetic factors distinguish these two classes. A case study of epidermal differentiation found that cohesin was enriched at "stable" chromatin interactions (Rubin et al., 2017), but it is still unknown what factors cause the preferential loading (or removal) of cohesin at different sites.

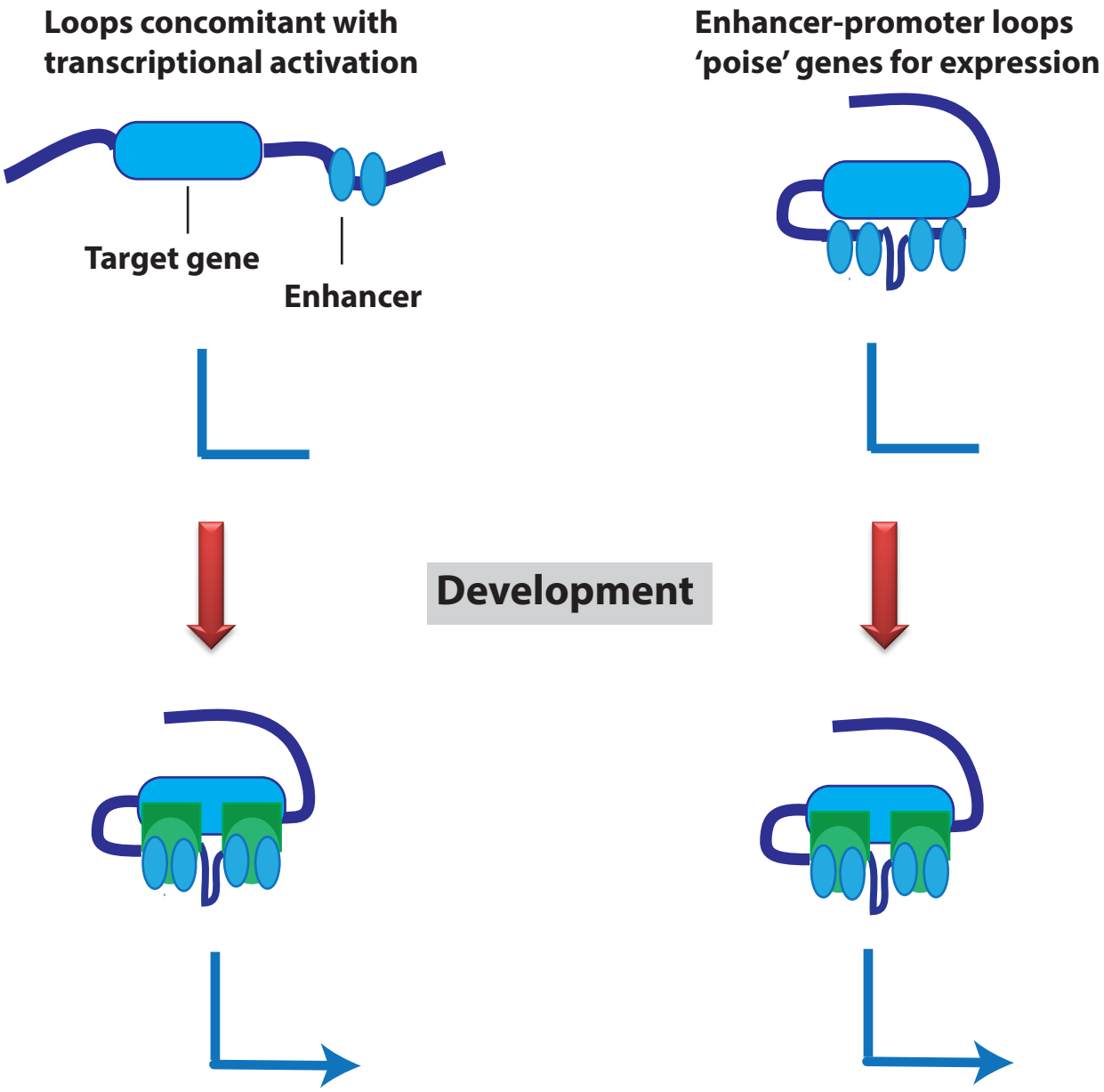


Fig 1. Chromatin loops as stable or dynamic structures.
Two models of chromatin loop dynamics during cell development. Left: target gene promoter is brought into proximity with enhancer at onset of transcriptional activation. Right: chromatin loop precedes transcriptional activation. The factors promoting transcription (green) may be brought concomitantly (left) or after (right) chromatin looping.

2.2 Topological associated domains (TADs) - units of genome folding

At the kilobase-to-megabase scale, genome-wide 3C (Hi-C; see section II - 2.4) studies have revealed that metazoan genomes are organized into discretely folded modules, termed topologically associated domains (TADs), whereby genomic interactions are strong within the domain but are sharply reduced on crossing a boundary between two TADs (Dixon et al., 2012a; Sexton et al., 2012a). TAD organization correlates well with histone modifications, coordinated gene expression, lamina association, and DNA replication timing, and their borders are enriched with binding sites for insulator proteins (Dixon et al., 2012a; Le Dily et al., 2014; Nora et al., 2012a; Pope et al., 2014), suggesting that they may represent functionally autonomous units of the genome. In support of this, TADs appear to delimit the functional range of enhancer activity (Symmons et al., 2014); naturally occurring TAD border deletions have been shown to permit aberrant enhancer-promoter contacts with concomitant developmental defects (**Fig 2**) (Lupiáñez et al., 2015). Further, pathological genomic duplications have been shown to not cause a phenotype if the duplicated region is insulated from the surrounding genes by forming a completely new TAD (Franke et al., 2016).

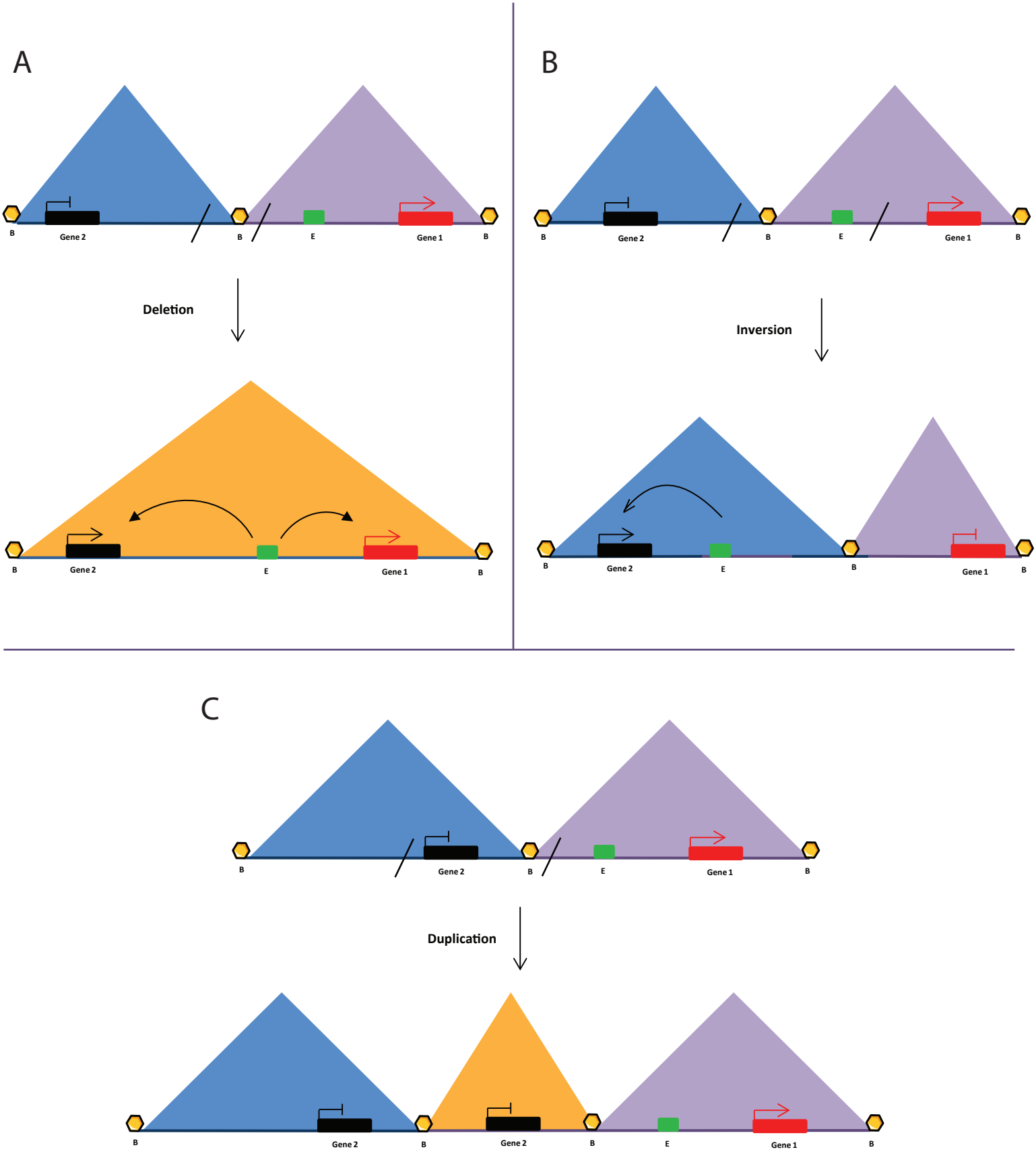


Fig 2: TADs may define gene regulatory zones
A) TAD border deletion leads to aberrant enhancer-promoter contacts.
B) TAD inversion disrupts certain enhancer-promoter contacts and leads to aberrant contacts with other genes.
C) TAD border duplication creates new TADs which are functionally separate from neighboring regions.

2.2.1 How are TADs formed and maintained?

Although TADs have been recently well studied, it remains unclear how they are formed or maintained. In fact, TAD borders are enriched in active genes and active histone marks such as RNA polymerase and H3K4me3 as well as "architectural" proteins cohesin and CTCF (Dixon et al., 2012b; Sexton et al., 2012b). As these factors are also enriched in chromatin loops, TADs could be a consequence of very strong interactions between TAD borders (Rao et al., 2014a). However, many TAD borders do not contain CTCF or cohesin and importantly the large majority of binding sites are not TAD borders. Deletions of single CTCF sites cause mild effects on the overall TAD structure but they may have important functional consequences by aberrant enhancer promoter communications (V. Narendra et al., 2015; Varun Narendra, Bulajić, Dekker, Mazzoni, & Reinberg, 2016). Interestingly, a very recent study with a complete ablation of CTCF in pluripotent cells caused a severe disruption of TADs (~80% of TADs disappeared) with a genome wide misregulation of gene transcription (Nora et al., 2017). Further, a total and systematic TAD loss has been observed with a complete ablation of cohesin (Rao et al., 2017; Wutz et al., 2017). Therefore, cohesin plays an essential role for TAD formation and maintenance, whereas CTCF is complementary to TAD stabilization. To date the best model to explain TADs and these phenotypes is the loop extrusion model, which gives a rationale for the observations of relatively uniform intra-TAD interactions, and the prevalence for convergent CTCF elements at their borders (Alipour & Marko, 2012; Fudenberg et al., 2016; Sanborn et al., 2015). Outlined in **Fig 3**, the model entails (i.) binding of an extrusion factor (or factors) at random positions in the genome; (ii.) physical extrusion of a chromatin loop, starting from this bound site, by two components of the extrusion factor translocating in opposite directions; (iii.) growing of the extruded loop, with a physical equilibrium between extrusion and disassociation of the extruding factor; (iv.) barriers to extrusion at specific regions within the genome, such as TAD borders. As

extrusion occurs by bidirectional translocation of the chromatin fiber, asymmetric barrier elements would need to be in a convergent orientation to function as TAD borders. CTCF sites thus fit in the model as candidate barrier elements to loop extrusion. Cohesin is the primary candidate for the extrusion factor, based on what is known about how the ring structure can organize tethered sister chromatids (K Nasmyth, 2001). The frequent co-occupancy of CTCF and cohesin at TAD borders could thus be interpreted as stalled loop extrusion complexes, which are more stable than actively translocating regions and are thus more frequently detected in chromatin immunoprecipitation studies. A prediction of the loop extrusion model is that the residence time of the extruding factor would determine the loop/domain size. In further support for cohesin playing this role, deletions of the cohesin loading factors SCC2 (Nipbl)/SCC4 or release factor, WAPL, in a human haploid cell line reduce or increase the average chromatin loop size, respectively (Haarhuis et al., 2017). Similar findings with depleted cohesin unloaders have been independently reported (Wutz et al., 2017). Interestingly, TAD structures were weakened but not completely destroyed in these studies, in contrast to the extreme effects of deleting the Nipbl cohesin loader in mouse liver (Schwarzer et al., 2017), suggesting that cohesin may sometimes be inefficiently loaded and unloaded from interphase chromosomes in the absence of these factors, and/or that extruding factors other than cohesin can also be present. It is currently unclear where cohesin-mediated enhancer-promoter interactions fit into this model. Large transcription complexes with RNA polymerase and its co-activators could reasonably be a barrier to loop extrusion, potentially explaining why active genes are frequently found at TAD borders (Dixon et al., 2012a; Nora et al., 2012a). Enhancer-promoter loops could thus conceivably be a metastable loop extrusion intermediate. However, cohesin binding is not detected at many chromatin interactions (e.g. (Rubin et al., 2017)), so loop extrusion may be mediated by other factors or not required at loops stabilized by multiple protein-protein interactions.

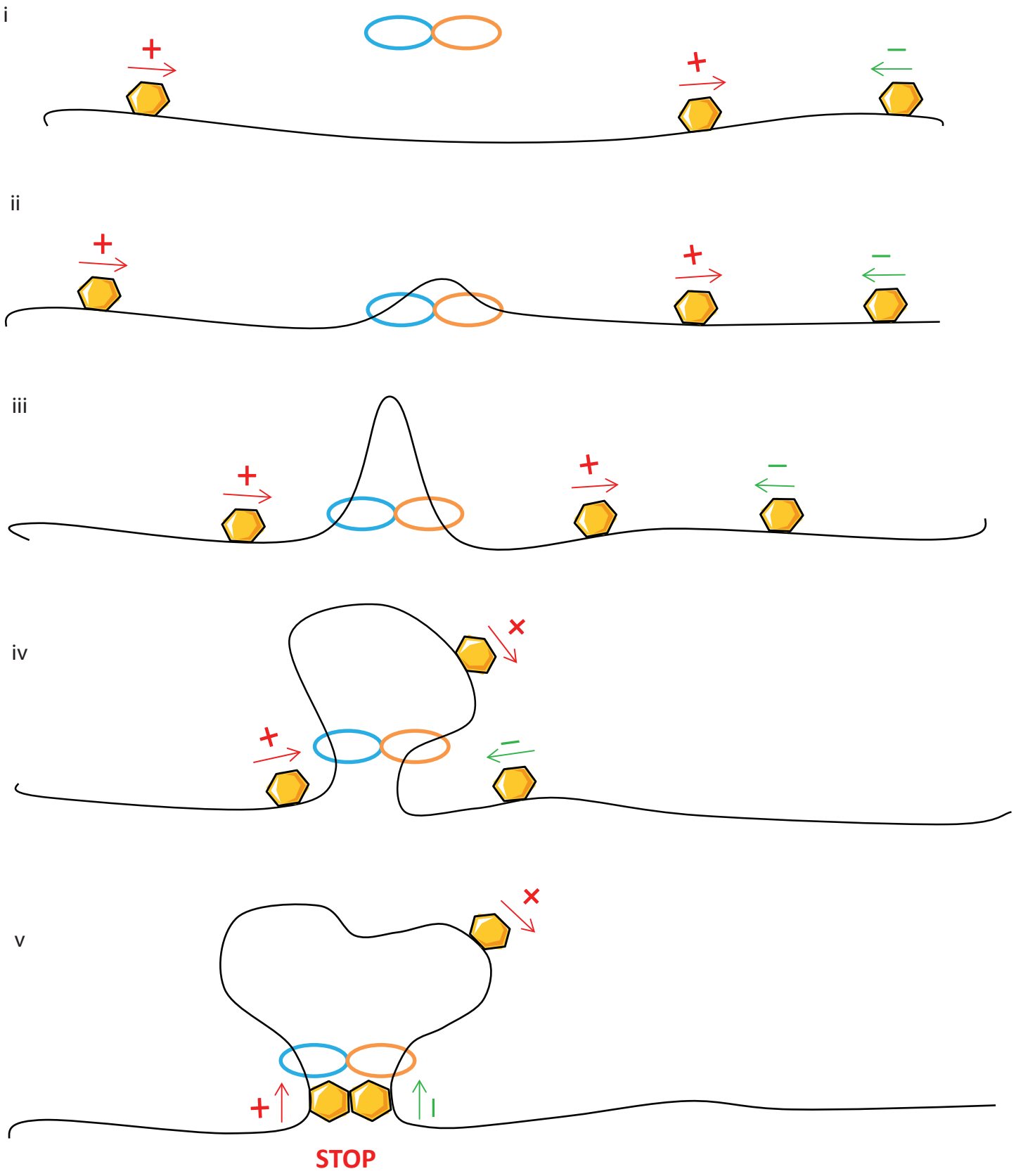


Fig 3: The loop extrusion model for TAD formation.
An extrusion factor (blue and orange ovals), often proposed to be cohesin, binds to chromatin and extrudes a growing loop by translocating in opposite directions. The growing loop is stalled when the base contains two barrier elements (yellow hexagons) in convergent orientation, proposed to be CTCF sites. The equilibrium of growing, stalled and disassembled extruded loops may explain TAD organization and the dependence of these structures on cohesin and CTCF.

2.2.2 TAD dynamics during development

Hi-C studies in disparate cell lines and tissues found that most TADs are invariant with cell type (Dixon et al., 2012a, 2015). This structural conservation is further supported by DNA fluorescent *in situ* hybridization (FISH) experiments coupled to super-resolution microscopy (Fabre, Benke, Manley, & Duboule, 2015), and even applies to syntenic chromosomal regions within different species (Pope et al., 2014; Vietri Rudan et al., 2015). Considering that TADs correlate so well with epigenetic marks, which are themselves extremely developmentally plastic (Roadmap Epigenomics Consortium et al., 2015), this observation was initially surprising. Higher-resolution studies at selected loci during pluripotent cells differentiation identified remodeled enhancer-promoter interactions as well as "sub-TADs" within larger conserved domains (Phillips-Cremins et al., 2013a). It was thus proposed that finer scale chromatin topology dynamics accompany cell reprogramming within a stable larger-scale architecture defined by TAD organization.

Original studies dedicated to investigate the spatial and temporal collinearity of mouse Hox gene expression, have demonstrated that the Hox loci form distinct topological domains with the active domain expanding and the silent domain shrinking according to collinear gene activation (Noordermeer et al., 2011). In fact, the mouse Hox genes are sequentially activated during development and according to anterior-posterior body position, in the order of the genes along the chromosome fiber. Hox gene expression is accompanied by a loss of H3K27me3-coated chromatin and concomitant gain of active histone marks(Soshnikova & Duboule, 2009). Developmentally dynamic "sub-TADs" within larger, stable domains have also been reported (Phillips-Cremins et al., 2013a). Current Hi-C data can support two conflicting models of TAD dynamics. First, TADs represent "ground state" chromosomal folding, on which other regulatory mechanisms (e.g. histone modifications, specific regulatory chromatin loops) are overlaid. Alternatively, TAD organization is responsive to

underlying chromatin state, but the limited resolution of current Hi-C studies overlooked subtle tissue-specific differences in limited resolution of current Hi-C studies overlooked subtle tissue-specific differences in chromosome conformation. The most high-resolution Hi-C experiment to date, performed on mouse ES cells during neuronal differentiation, has blurred these two models, reporting many developmentally stable and dynamic TADs (Bonev et al., 2017). As for promoter-enhancer interactions, it is currently unclear what features distinguish a remodeled TAD from a stable one; remodeled TADs are associated with transcriptional induction, but many stable TADs present similar expression differences, and ectopic induction of otherwise remodeled genes is insufficient to alter their TAD architecture (Bonev et al., 2017).

In summary, TADs are mostly evolutionarily and developmentally stable structures, organizing seemingly autonomous regulatory domains in pluripotent and differentiated cells. We have made much progress in understanding their basic architectural principles, which in turn may ensure specific and efficient homing of genes to their regulatory elements. However, we have much to learn about the "fine print" of these architectural "blueprints", in particular if and how specific intra-TAD interactions are set up, and whether they contribute to TAD stability (Giorgetti et al., 2014). Furthermore, much remains to be learned on the interplay between "stable" bulk TAD organization and "dynamic" chromatin ultrastructure during the large-scale transcriptional changes accompanying development.

2.3 Genomic compartments

Individual chromosomes are spatially organized into large compartments. These were first inferred from the "plaid" patterns of Hi-C contact maps, suggesting that multi-megabase regions are organized into one of two categories, "A" or "B," whereby preferential

interactions occur between regions belonging to the same category, with very little mixing of the resulting A and B compartments (Lieberman-Aiden et al., 2009a). Epigenomic profiling of these compartments revealed that "A" chromatin is generally "open" and transcribed, whereas "B" chromatin carries repressive histone modifications and is more gene-poor. More refined analyses on higher-resolution Hi-C datasets are able to further split the A and B compartments into subcategories of preferentially interacting regions, based on location relative to the centromere(Yaffe & Tanay, 2011) or more specific histone modifications (Rao et al., 2014b). This compartmentalized organization could be a general result of preferential homotypic interactions between genomic elements sharing the same functions and chromatin states, as has been observed in the clustering of co-transcribed genes (Schoenfelder et al., 2010) or genes repressed by Polycomb (Bantignies et al., 2011). Self-organization models propose that this chromatin compartmentalization allows robust genomic control by ensuring that co-expressed genes share access to the same regulatory factors (Sexton & Cavalli, 2015). Such a model is difficult to experimentally assess, although abrogation of one gene has been shown to perturb expression of distal interacting genes (Bantignies et al., 2011; Fanucchi, Shibayama, Burd, Weinberg, & Mhlanga, 2013).

Until recently, it was unclear whether compartments were essentially larger-scale TADs, subject to the same organizational principles, considering that TADs form sub-domains (Phillips-Cremins et al., 2013b) and the patterns of inter-TAD interactions in Drosophila essentially follow genome compartments (Sexton et al., 2012b). However, the

recent perturbation studies in different cell types demonstrated a clear decoupling of TAD and compartment organization in mammals: CTCF ablation disrupted TADs with minimal effects on compartments (Nora et al., 2017), whereas cohesin ablation actually *reinforced* compartmentalization (Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017). This suggests that not only do TADs and compartments arise by different mechanisms, but also that they may be competing processes organizing chromosome folding. This concept was taken further by re-analysis of high-resolution Hi-C datasets in multiple eukaryotic species, which re-classified many TAD borders containing active genes as small A compartments which break up the B compartments within which they reside (Rowley et al., 2017). One model to explain competition between TADs and compartments is that the latter are set up by self-organization principles, allowing *general* reinforcement of regulation of entire programs of genes. However, this creates a search space too large for the efficient homing of enhancers to *specific* gene targets. TAD organization restricts this search space predominantly to the extruded loop domain, thus increasing transcriptional fidelity, and explaining why perturbations reducing TAD "insulation" has widespread but not complete positive and negative effects on gene expression (Nora et al., 2017; Sofueva et al., 2013; Zuin et al., 2014).

2.4 Chromosome territories

At the coarsest level, interphase chromosomes occupy distinct regions within the nucleus, termed chromosome territories, which can be discerned by light microscopy after FISH with cocktails of labeled probes (Cremer & Cremer, 2001). Hi-C studies also reveal chromosome territories, based on the finding that inter-chromosomal contacts are generally less frequent than interactions between the most distal regions of the same chromosome (Lieberman-Aiden et al., 2009a; Sexton et al., 2012a). Comparing the frequencies of different chromosome pair interactions also supports previous FISH studies suggesting that chromosome territories have preferential partners within the nuclear space (Boyle et al., 2001; Parada, McQueen, & Misteli, 2004). Despite reports of specific functional *trans* interactions in mammalian cells, based on FISH and 3C/4C experiments (Clowney et al., 2012; Schoenfelder et al., 2010), these are not readily detected in Hi-C experiments. It is not clear whether such inter-chromosomal interactions are restricted to very specific cell types, or are too infrequent to be robustly detected above background in genome wide studies.

II: Assessing chromatin interactions

Over the past decades, genome organization has been assessed using two major approaches: microscopy, often based on FISH and molecular biology approaches, mostly variants of 3C. As it is not the major scope of my thesis, I will give a brief summary of the state of the art of microscopy approaches to study chromatin interactions, before giving a more in-depth appraisal of the different 3C variants that are used or discussed in this study.

1. Microscopic approaches

FISH is a cytogenetic technique which uses fluorescent probes to target specific loci or even whole parts of the chromosome in fixed cells. The first use of this technique was to identify the position of ribosomal DNA within the nucleus of a frog egg (Gall & Pardue, 1969). The major advantages of this approach over 3C methods is that it is a single-cell technique, allowing cell-to-cell heterogeneity to be assessed, it enables true measurements of distances between genomic loci, it can be directly coupled with immunolabeling of nuclear landmarks to give different information on nuclear organization, and it is not restricted to pairwise interactions. The major limitations are that it is unfeasible to perform FISH experiments at a genome-wide scale throughput, the resolution is limited, and the dynamics of chromosome interactions can still not be addressed. However, recent advances are beginning to address some of these limitations.

An array of super-resolution light microscopy methods (Sydor, Czymmek, Puchner, & Mennella, 2015) has allowed the diffraction limit of light to be overcome to visualize

structures at a precision of several nanometers. Based on its applicability to standard FISH sample preparation techniques and fluorophores, structured illumination microscopy (SIM) is the most common technique to explore chromatin interactions, doubling the effective spatial resolution by using interference-generated light patterns. DNA FISH coupled to SIM has been used to study the structures of TADs (Fabre, Benke, Joye, et al., 2015; Nora et al., 2012b) and individual gene loci (Patel et al., 2013; van de Corput et al., 2012). More recently, the incorporation of photoswitchable dyes into FISH probes has allowed the technique to be coupled to stochastic optical reconstruction microscopy (STORM), obtaining "structures" of labeled chromatin regions at ~50 nm resolution (Beliveau et al., 2015; Fabre, Benke, Joye, et al., 2015). Various throughput bottlenecks in the FISH technique have also been addressed. Automated image analysis tools allow FISH to be performed at the throughput of large-scale screens (Shachar, Voss, Pegoraro, Sciascia, & Misteli, 2015) and innovations in synthetic oligonucleotide probe design allow thousands of probes to be simultaneously used in one experiment (e.g., (Beliveau et al., 2015)). For instance, this Oligopaint technology was recently used to simultaneously label all individual TADs on one human chromosome, with ~1000 probes per TAD, and supported the discrete folding of TADs into A/B compartments within single cells (Wang et al., 2016). To date, no study has been published combining all of these innovations to provide high-resolution, high-throughput FISH screens, but this is feasible in principle. These first studies have focused on tiling approaches to label TADs or

loci with multiple probes and elucidate structure; it will be interesting to see how super-resolution approaches perform in assaying specific chromatin looping interactions.

To assess the dynamics of chromatin architectures, live microscopy after labeling specific loci is required. Until recently, the major means of labeling chromatin for such experiments was to insert multiple copies of bacterial repressor sequences, such as lac or Tet, which are bound by fluorescently tagged repressor proteins (Robinett et al., 1996). For example, the Tet system has been used to show the spatial constraint of gene loci within constrained domains, presumably TADs (Lucas, Zhang, Dudko, & Murre, 2014). However, this approach is limited for various reasons. A lack of orthogonal systems means that multiple-labeling experiments are extremely difficult, and the requirement of large copy numbers of repetitive sequences for a robust fluorescent signal makes genetic manipulation very difficult. Furthermore, the insertion of ~10 kb of ectopic sequence is likely to affect the local chromatin topology of the locus of interest; the lac repressor, for instance, has been shown to induce local chromatin silencing in yeast (Dubarry, Loïodice, Chen, Thermes, & Taddei, 2011). Recent developments have overcome or reduced these shortcomings and give much promise for future experiments assessing the dynamics of chromatin topologies. One method, termed ANCHOR, also uses ectopic bacterial DNA sequence/fluorescently tagged binding protein combinations, in this case the *parS*/ParB system for plasmid segregation (Saad et al., 2014). However, unlike for lac or Tet repressors, ParB has self-oligomerization properties, allowing robust signals to be obtained for small parS copy numbers (~1 kb in total), and the ANCHOR system has been shown to have minimal effects on endogenous transcription when inserted into specific loci in yeast (Saad et al., 2014). At least four ANCHOR orthologs have been developed (K. Bystricky, personal communication), giving great promise for multicolor experiments to assess chromatin interaction dynamics, especially since the recent revolution in genome editing tools allows for the specific insertion of *parS*

sequences into almost any locus. For example, ANCHOR was recently used to assess dynamics of a gene after acute induction with estradiol (Germier et al., 2017).

The other major approach to visualize genomic loci *in vivo* utilizes the aforementioned genome editing tools directly, fluorescently tagging the nuclease-dead variants of TALE proteins or CRISPR/Cas9 constructs that are engineered to bind specific DNA sequences (B. Chen et al., 2013; Ma et al., 2016; Miyanari, Ziegler-Birling, & Torres-Padilla, 2013). The greatest advantage of this approach is that endogenous loci can be directly visualized, with no need to insert ectopic sequences. However, the signal strength from single TALE or CRISPR/Cas9 binding sites is insufficient for robust visualization; published applications of this method are mostly restricted to labeling repetitive DNA sequences, such as satellites or telomeres. One case where a single-copy locus was visualized with GFP-tagged CRISPR/Cas9 required 30 guide RNAs, tiled over a 2 kb element (B. Chen et al., 2013), suggesting that it will be very challenging to apply this to most genomic locations. However, recent studies are reporting further successes with the technique(Gu et al., 2018). Furthermore, both TALE and CRISPR/Cas9 approaches face challenges in being used for multi-label experiments. The use of TALEs is in principle only limited by the number of available fluorophores, but it is laborious to redesign and produce a new TALE for each locus of interest. Until very recently, CRISPR/ Cas9 was limited to single-label experiments, since the different guide RNAs recruit the same tagged CRISPR protein. However, multiple labeling is now possible, either through the use of orthologous CRISPR systems from different bacterial species or adding different stem loops to the guide RNA scaffolds, which in turn recruit different tagged binding proteins to the complex (Ma et al., 2016). Further technological advances in microscopy, CRISPR applications, and ANCHOR are likely to open up a new frontier where chromatin architecture dynamics can be fully addressed.

2. Chromosome conformation capture and its variants

2.1 3C

In addition to direct visualization by microscopy, chromosome structure can be deduced based on the frequencies with which genomic segments contact each other within a cell population. The chromosome conformation capture (3C) method allows for the detection of such specific pairwise interactions (Dekker, Rippe, Dekker, & Kleckner, 2002) . Briefly, cells are first fixed with formaldehyde to create covalent bonds between chromatin fibers that are in sufficient physical proximity *in vivo* during the cross-linking process. The chromatin is then digested with a restriction enzyme and re-ligated to form chimeric products between such crosslinked restriction fragments, irrespective of their separation on the linear chromosome fiber. Specific interactions are subsequently assessed by quantitative polymerase chain reaction (PCR) with primers designed to candidate genomic regions. Basic polymer physics, supported by light microscopy studies (Mateos-Langerak et al., 2009), suggest that the probability of an interaction between two chromatin regions decreases rapidly (on a power scale) with increasing genomic separation between them. Using appropriate controls(Dekker, 2006), 3C can identify specific chromatin looping events, whereby the interaction between two distal elements is stronger than with intervening regions. The most frequently described chromatin loops in the literature are those between promoters and distal enhancers (see section I), but 3C has also identified other classes of chromatin looping events with potential functional significance. These include contacts between promoters and gene terminators (N. Le May,

Fradin, Iltis, Bougnères, & Egly, 2012; Tan-Wong et al., 2012), insulator-mediated loops(Kurukuti et al., 2006; Splinter et al., 2006), and topologies linked to recombination events (L. Chen, Carico, Shih, & Krangel, 2015). Despite its low throughput, 3C results have also been used to infer physical models of chromosome folding (Court et al., 2011; Dekker et al., 2002).

2.2 4C and 5C

Various derivatives of the "one-to-one" 3C method have benefited from the recent explosion in high-throughput sequencing; instead of relying on PCR amplification from specific primers, 3C ligation products can be more globally detected for systematic mapping of chromatin interactions. Briefly, 4C (circular 3C) is a "one-to-all" method allowing all interactions with one specific bait region to be assessed, first by hybridization to microarrays (Simonis et al., 2006) , and then by direct sequencing (van de Werken et al., 2012), which has been used to identify enhancer-promoter interactions at high resolution (Ghavi-Helm et al., 2014), to assess specific spatial chromatin domains such as TADs (Lupiáñez et al., 2015),(Noordermeer et al., 2011), and to identify networks of gene co-associations (de Wit et al., 2013; Schoenfelder et al., 2010). Aside from being restricted to just one bait, a limitation of conventional 4C-seq is that each interacting restriction fragment can only produce one specific PCR product; quantitative analysis is confounded since PCR duplicates cannot be distinguished from biological signal. This is partially overcome by sub-sampling and

assessing interactions as sliding windows of multiple fragments (de Wit et al., 2015a), but UMI (unique molecular identifier)-4C has also recently been developed to completely remove PCR duplicates (Schwartzman et al., 2016). 5C (3C carbon copy) is a "many-to-many" method using large sets of multiplexed primers to simultaneously assess thousands of chromatin interactions (Dostie et al., 2006) and has been used to assess promoter interaction landscapes (Sanyal et al., 2012) and the structures of specific chromosome domains (Nora et al., 2012a). This technique relies on the primers hybridizing exactly to ligation junctions for subsequent pair ligation, and appears to suffer from large technical biases in oligonucleotide hybridization efficiency, more so than other methods (see Capture-HiC, **2.2.5**).

2.3 Hi-C

As sequencing throughput increases, it has become feasible to globally assess all chromatin interactions within a population ("all-to-all" methods) simply by sequencing 3C ligation products. This pioneering technique, termed Hi-C, was first developed in human cell lines (Lieberman-Aiden et al., 2009a), which also included the innovation of introducing biotin to 3C ligation junctions, allowing them to be purified before sequencing. Hi-C has subsequently been used to derive chromatin interaction maps for a large number of species (Vietri Rudan et al., 2015) (see overview in (Ben Zouari *et al.*, 2017)). These landmark "interactome" maps have allowed chromatin architectural principles inferred from case studies to be generalized to eukaryotic genomes and have further uncovered the previously described functional principles of chromosome folding (**Fig 4**). In principle, Hi-C can resolve interactions to the level of individual restriction fragments. However, its strength in assessing all possible chromatin interactions is also one of its major disadvantages: the numbers of possible ligation products that can be detected is much greater than current sequencing output. For example, the mouse

genome consists of ~1.5 million fragment ends after digestion with a restriction enzyme commonly used in Hi-C, giving ~$1x10^{12}$ possible pairwise combinations of ligation products. A naive "1x" coverage of this interaction space thus requires at least 1500 lanes of the current standard high-throughput sequencer. In order to get robust read counts, Hi-C data are usually assessed over bins of multiple pooled restriction fragments, lowering the resolution of called interactions. As Hi-C datasets with increasing sequence depths have been produced, specific looping events have been successfully resolved ((Bonev et al., 2017; Rao et al., 2014a); **Fig 4**), but extrapolation of the numbers of loops identified by 4C experiments to the entire genome suggests that only a small subset of the strongest interactions have been found.

Despite new computational methods attempting to enhance Hi-C resolution (Grubert et al., 2015), a better approach may be to reduce the complexity of the pool of sequenced ligation products. This allows an equivalent number of reads to give higher-resolution interaction information, albeit for a more limited subset of the possible genomic space. 4C is the most extreme case of this approach, whereby only 100,000 to 2 million reads are required for the comprehensive interactome of one specific bait (Schwartzman et al., 2016; van de Werken et al., 2012). Other variants, described below, aim for a compromise between genomic coverage and resolution.

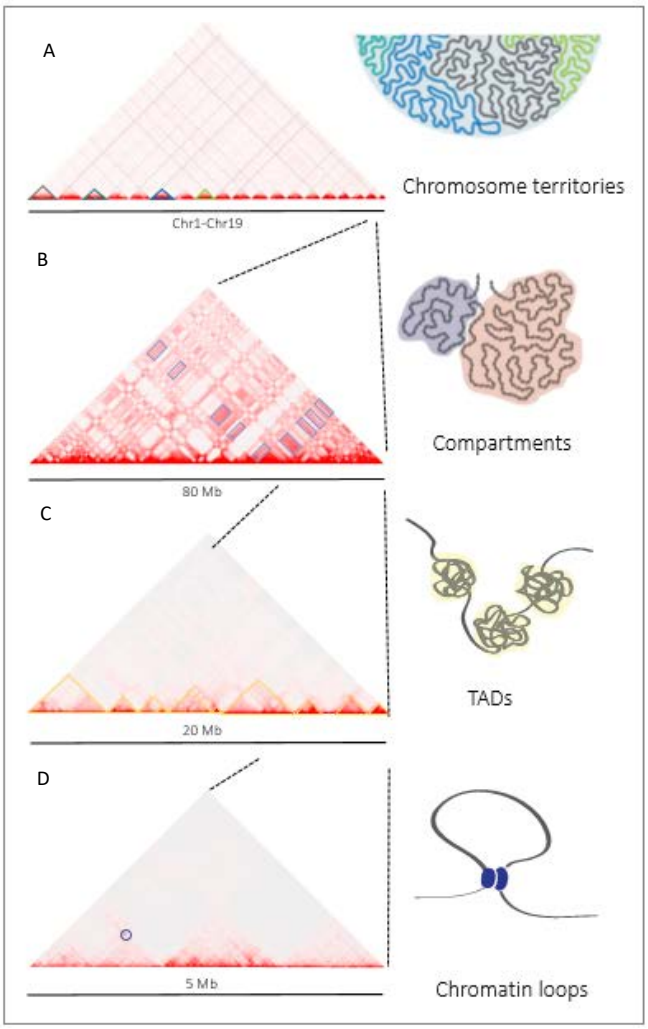


Fig 4: Inferring chromatin architectures from Hi-C contact maps.

Deeply sequenced Hi-C data from mouse double-negative (DN3) thymocytes were generated in our group (see Results) and are presented as two-dimensional contact heat maps (left), showing the numbers of reads measured for pairwise combinations of genomic loci. For example, the interaction highlighted by a blue circle in (D) represents frequent contacts between the genomic loci denoted by asterisks. Features of note are outlined in blue. (A) The strongest interactions are constrained within individual chromosomes, supporting spatial organization of the genome into discrete chromosome territories. (B) The plaid pattern of alternating strong and weak interactions on the heat map indicates compartmentalization of the genome, whereby certain genomic regions preferentially co-associate, and others prefer to be segregated. (C) The triangular patterns close to the heat map baseline indicates discretely folded chromosomal domains (TADs). Note that the TAD structure can be quite complex, with "sub-TADs" within larger domains. (D) For sufficiently deeply sequenced datasets, Hi-C maps can reveal punctate specific interactions, suggestive of chromatin loops.

2.4 ChIA-PET and HiChIP

ChIA-PET (chromatin interaction assessment with paired-end tags) essentially couples chromatin immunoprecipitation (ChIP) with Hi-C (with some technical differences) to target interactions between genomic regions bound to a specific protein (Fullwood et al., 2009). This technique has been used to identify the networks associated with CTCF (Handoko et al., 2011), RNA polymerase II (G. Li et al., 2012) , and the estrogen receptor (Fullwood et al., 2009). Very recently, a variant called HiChIP was developed, which appears more efficient and can be used on smaller populations of cells (Mumbach et al., 2016). Notably, HiChIP was coupled with immunoprecipitation for H3K27ac to catalog promoter-enhancer contacts in rare human T cell subtypes (Mumbach et al., 2017). However, a major limitation of these approaches is that the relative numbers of ligation products that are sequenced depend on the complex combination of chromatin interaction frequency and immunoprecipitation efficiency (i.e. how strongly the region is bound by the assessed protein), which cannot be distinguished. Inappropriate importance may thus be given to a very weak interaction between two regions which strongly bind their factors. This problem is even more serious when trying to compare interactomes in two cell types, which have very different ChIP-seq profiles.

2.5 Capture-HiC (CHi-C)

Several groups have recently reduced the complexity of their sequenced Hi-C material by first capturing the material on libraries of thousands of oligonucleotides complementary to selected restriction fragments (Hughes et al., 2014; Joshi et al., 2015; Kolovos et al., 2014; Sahlén et al., 2015; Schoenfelder, Furlan-Magaril, et al., 2015). Unlike 5C, these oligonucleotides are

not constrained to the exact ligation junction, allowing some flexibility in their design and reduction of perceived technical problems (e.g. controlling better for GC-content and unique sequence). Unlike ChIA-PET, any probe-to-probe differences in capture efficiency can be expected to be the same for different cell types (see Results for confirmation), allowing fairer comparisons of their interactome. To date, the majority of published Capture-HiC (CHi-C) studies have used single (or limited) dispersed probes to generate highly multiplexed 4C-like data for hundreds or thousands of *cis*-regulatory sequences, predominantly promoters ((Schoenfelder, Furlan-Magaril, et al., 2015)·(Freire-Pritchett et al., 2017; Hughes et al., 2014; Mifsud et al., 2015; Sahlén et al., 2015)·(Javierre et al., 2016)), but also DNase-hypersensitive sites (Joshi et al., 2015) and targeted DNA break regions (Aymard et al., 2017). However, designs of tiled oligonucleotides spanning a contiguous region of interest can generate 5C-like data for high-resolution assessment of TAD structures (Franke et al., 2016; Kolovos et al., 2014). Very recently, a flurry of promoter CHi-C studies was published, which largely focused on promoter-enhancer interactions in different cell types, many addressing the same questions that I had posed during my thesis (see Research Aims). The overall conclusions that could be made from these studies are described below:

Firstly, whereas various epigenetic signatures such as H3K27ac can predict enhancer activity, it is known that they frequently do not regulate their nearest gene on the chromosome fiber (Sanyal et al., 2012). H3K27ac regions which interact with promoters, as measured by CHi-C, appears to be a reliable predictor of enhancers regulating that particular gene (Mifsud et al., 2015; Schoenfelder, Furlan-Magaril, et al., 2015). Secondly, previously cryptic sequence variants within intergenic elements that are linked to disease can be better understood by identifying which genes they interact with, presumably acting as cryptic or deregulated enhancers (Mifsud et al., 2015). Third, when comparing different cell types, those which are closest in developmental stage have the most similar enhancer-promoter

interactomes, consistent with a more similar "epigenetic state" (Javierre et al., 2016). Fourth, when comparing cells across a differentiation pathway, enhancer interactions can be very dynamic/cell-type-specific (Siersbæk et al., 2017), although more developmentally stable interactions are often also found (Freire-Pritchett et al., 2017; Rubin et al., 2017), supporting both the instructive and permissive models of enhancer-promoter looping in transcriptional control. To now, it is unclear exactly what differentiates instructive and permissive loops. In one study of acute (4 hr) adipocyte differentiation, which reported predominantly instructive loops, the concomitant looping and gene expression also correlated with gain in H3K27ac at the enhancer (Siersbæk et al., 2017). However, H3K27ac is observed in both instructive and permissive loops in another study of epidermal differentiation over days (Rubin et al., 2017); instead, they reported that cohesin appeared to be specific to the stable loops, and specific epidermal transcription factors correlate with certain (but far from all) gained enhancer interactions. It is unclear if cohesin plays a direct role in transcriptional poising or firing, or if it is just easier to detect on more "stable" chromatin loops.

III: Thymocyte development

The adaptive immune system requires the efficient recognition of essentially any "non-self" antigen derived from an infectious agent. In mammals, this is achieved by antibodies or immunoglobulins derived from B (bone marrow-derived) cells, and paralogous receptors on T (thymus-derived) cells circulating around the body. In both cases, the tremendous diversity of the receptors is driven by recombination events between different variable cassettes at the immunoglobulin or T cell receptor (TCR) gene loci in developing lymphocytes. Errors in this process, or of the subsequent selection for productive rearrangements and removal of rearrangements causing recognition of "self" antigens, are linked to numerous diseases, including leukemia, immunodeficiencies or autoimmune disorders. As a model system for epigenetic and chromatin topology changes during development, our group study mouse thymocyte maturation.

1. Thymopoesis

The differentiation of T cells goes through subsequent steps, starting from arrival of progenitor cells at the thymus. The early thymic progenitors (ETP) lose gradually their multipotency which involves a chemokine receptor CCR9 and a ligand for P-selectin expressed on the thymic epithelium. This process takes place in double negative thymocytes (DN) which do not express CD4 or CD8 receptor (Bell & Bhandoola, 2008). The DN thymocytes are themselves divided into different subsets (DN1 to DN4) based on the expression of two receptors CD44 and CD25 (Schlenner & Rodewald, 2010). The loss of multipotency is not complete before the DN2 stage in which different genes (e.g. Notch1) and transcription factors (e.g. Runx1, GATA-3 and others) cooperate to initiate T cell differentiation. The most important rearrangements of variable gene segments occur in DN3 cells (Schlenner & Rodewald, 2010). Indeed, the first checkpoint for TCR gene

rearrangement (β-checkpoint) is at DN3 stage. The TCR gene encodes for TCR receptor expressed in the surface of T cells which is responsible for foreign antigen recognition (Buer, Aifantis, DiSanto, Fehling, & von Boehmer, 1997). At the β-checkpoint, only the thymocytes expressing T cell receptor which are capable of binding to major histocompatibility complex (MHC) are maintained (Buer et al., 1997). Subsequently, the thymocytes proliferate and become double positive cells (DP), expressing both CD4 and CD8. The maturation of these double positive T cells undergoes further lineage commitment to single positive (CD4+ "helper" or CD8+ "cytotoxic") T cells. Due to the biological and medical importance of the β-checkpoint, a large body of work has characterized the transcriptome and epigenome of mouse DN and DP cells (Egawa & Littman, 2011; Koch et al., 2011; Pekowska et al., 2011; J. A. Zhang, Mortazavi, Williams, Wold, & Rothenberg, 2012), providing an extremely useful reference for which to compare any developmental dynamics of chromatin topology. Hi-C studies have also been made in mouse thymocytes (G. Hu et al., 2018; Seitan et al., 2013), but not at a deep enough coverage to allow fine-scale chromatin architectural dynamics to be explored.

2. Transcription factors during Thymocyte differentiation

During T lineage differentiation, many T lineage regulatory factors are implicated, with no single master regulatory factor. Ikaros is very important in the earliest DN1 pro-T cell stage and controls the frequency with which lympho-myeloid precursors can embark on a lymphoid pathway (Ng, Yoshida, Zhang, & Georgopoulos, 2009). Not limited to early T cell stage, Ikaros is also required to regulate the Notch signaling pathway in later stages (Tinsley et al., 2013). Other factors have been shown rather to block thymocyte maturation when over expressed, like GATA-3(David-Fung et al., 2009), with evidence showing that it drives pro-T cells to a distinctive form of lineage diversion (Scripture-Adams et al., 2014). In fact, GATA3 seems to be important for Th1/Th2 cell fate decision, where it acts by binding to distal

enhancers at immune regulatory genes in effector T cells (Kanhere et al., 2012). Runx1 is a crucial factor for establishing the hematopoietic stem cell compartment (de Bruijn & Speck, 2004). Recently, mutational studies on Runx1 binding sites in the enhancer region of the TCR-β gene showed an essential role of Runx1 in the initiation phase of TCR-β expression but not necessarily for maintaining the enhancer activity at later developmental stages(de Bruijn & Speck, 2004).

AIMS

Research Aims

To better understand the role of genome organization in transcription regulation, we wished to address the following questions:

(i) How are chromatin configurations altered during transcriptional changes accompanying development?

(ii) Is chromatin topology important in controlling cell differentiation and development?

To answer the challenging questions proposed above, we have chosen thymocyte development as a biological model for studying chromatin topology during development. Thymocyte development is a perfect system in which the effect of fine-tuning of transcriptional output on genome organization can be seen. Specifically, we interrogated specific chromatin architectures in mouse $CD4^-$ $CD8^-$ $CD44^-$ $CD25^+$ (DN3) and $CD4^+$ $CD8^+$ (DP) thymocytes, representing stages just before and after β-selection. For these populations, it is easy to obtain the pure and homogeneous populations that are required for a successful Hi-C experiment. To optimize the resolution and throughput of these studies, we performed two different CHi-C strategies, assessing the developmental dynamics of chromosome conformation on different scales.

To study TADs, we used the frequently-cutting restriction enzyme DpnII (for maximal resolution) and used tiled probes to the ends of all sufficiently long DpnII fragments flanking either side of eight selected TAD borders (covering 600 kb in total for each border). These borders, defined from Hi-C in ES cells (the only mouse dataset with sufficient reads for robust border calling at the date when starting the project) (Dixon et al., 2012a), are within a few kilobases of T cell lineage-specific genes. Three borders are close (<20 kb) to genes which are significantly upregulated on DN3-to-DP transition (*Nfatc3*, *Bcl6* and *Rag1*) and three borders

are close to genes significantly downregulated after β-selection (*Il17rb*, *Pla2g4a* and *Cdh1*). These genomic regions may thus be expected to undergo reorganization during thymocyte differentiation. Two control borders in the capture design are very close to genes which are expressed in both cell types (*Cd3* and *Zap70*).

The second strategy uses interspersed capture probes designed to ~22,000 promoters, covering nearly all known genes. This simultaneously generates 4C-like datasets with promoters as baits, allowing the systematic identification of their interactions with distal regulatory elements in DN3 and DP cells. We used the same restriction enzyme, HindIII, and capture strategy as adopted previously in mouse ES cells (Schoenfelder, Furlan-Magaril, et al., 2015).

The CHi-C experiments were performed by other members of the group. I performed the vast majority of the computational analysis of the Hi-C and CHi-C datasets, and comparative studies with the publicly available transcriptomic and epigenomic profiles. The results presented in my thesis are split into five sections:

1. Quality control of the Hi-C and CHi-C datasets;
2. Development of PromoMaxima, a new robust method to identify looping interactions from CHi-C data (Ben Zouari et al., in preparation).
3. Analysis of the promoter CHi-C data, uncovering extensive dynamics of chromatin looping interactions, linked to both transcriptional activation *and* repression (Molitor*, Ben Zouari* et al., in preparation).
4. Comparative analysis of the different methods available to call TADs, as applied to Hi-C or CHi-C data.
5. Analysis of the TAD CHi-C data, identifying a developmental robustness of most TAD architectures, but uncovering a significant minority that can be directly remodeled by transcription (Chahar, Ben Zouari et al., in preparation).

RESULTS

I. Hi-C and CHi-C quality control

The analysis workflow of Hi-C or CHi-C data starts with common steps of data preprocessing, before calling chromatin structures. Here, I present the preprocessing steps of Hi-C and CHi-C data: alignment, filtering, normalization and construction of matrices. Finally, I introduce some statistics to consider for the quality of Hi-C/CHi-C libraries.

1. Hi-C data processing

1.1 Sequence trimming and alignment

All Hi-C data in this study was generated using Illumina paired-end sequencing with 50 bp read length. The pipeline below used for Hi-C analysis applies to all Hi-C, Promoter-Capture and TAD-capture data. **Table 1** resumes the number of sequencing reads reported for each experiment and tissue. To process Hi-C data we used a custom pipeline originally from our collaborators (Sexton et al., 2012b) which has been optimized for parallelized computation on a cluster. In many respects, this pipeline is highly similar to the HiCUP used in other studies (Wingett et al., 2015).

The pipeline begins by splitting each of the two fastq files into smaller blocks containing 1 million single end reads. Each block goes through trimming step before mapping. The DNA products of Hi-C are chimeric and contain different regions of the genome ligated together. We expect a forward read maps to one ligation fragment whereas the reverse maps to the other. However, some reads may contain the ligation junction. Such reads may cause difficulties when trying to map to the reference genome. In order to save these reads, we identify the ligation junctions within the sequenced regions and truncate sequence downstream of the restriction enzyme recognition site. ~20% of total reads were trimmed, depending on the restriction enzyme used (**Fig 1A**). After trimming, each chunk then is

mapped to the mouse genome (assembly mm9) using the Bowtie program with these parameters (-t -B 1 -a -m 1 --best --strata --chunkmbs 200). Unmapped or non-uniquely mapped reads were discarded. The alignment is performed in parallel on a server cluster of the IGBMC platform. For one lane of Hi-C sequencing, there are around 200 jobs and it takes on average ten minutes to align 1 million single end reads. Paired end mappers should be avoided as they make assumptions about the insert size which are false for a ligation product such as in Hi-C data. Each fastq file produces a SAM file sorted by read name. In total, we generate hundreds of SAM files that correspond to the first and second unmated read, which then are merged into a single paired end text file, in which we keep only the chromosome and position information of each read and its mate. Then, each chunk file of paired-end goes through a deduplication step in which we remove all PCR duplicates resulting from pairs with both mates at exactly the same location. This is accomplished using a perl script that splits first the file of mapped-paired reads into chunks containing around 1 M read pairs each. Then, for each 1 M read pairs, we look for any duplicates. All this is done in the same time which takes in total 20 minutes using parallelized jobs on the IGBMC cluster. On average, we filter around 2 to 5% of PCR duplicated reads, implying that the Hi-C libraries have an excellently maintained complexity during the limited numbers of PCR amplification cycles necessary in the method (**Fig 1A**)

1.2 Filtering non valid reads

After read mapping and pairing, we filter reads that may come from Hi-C artefacts (**Fig 1B**):

- Self-ligation: where reads map to a DNA fragment, ligated to itself to form a circularized DNA.

- No-ligation: Reads that map to a single restriction fragment where both or one of the ends maps exactly on the restriction enzyme cut site, not a random site as would be expected from sonication of the Hi-C DNA during library preparation.
- No-restriction: Reads that span several short restriction fragments from a contiguous DNA, so are more likely to be non-digested genomic DNA than *bona fide* Hi-C products.

For each restriction enzyme, we made a text file that contains the position of each restriction fragment end in the genome with a corresponding unique ID. Based on this file as reference, we use a binary search to identify for each read pair the corresponding restriction fragment ends that they map to. The final output is a text file that we call "mat table" that contains unique fragment end pairs with their corresponding number of Hi-C reads. Each line in this file corresponds to unique fragment end pairs with the following columns: fend1, fend2, reads_number.

This file is the basis of all downstream analysis for Hi-C, promoter and TAD CHi-C.

1.3 Hi-C library quality control

Prior to performing any CHi-C, we usually sequence the Hi-C library in order to be sure of the library quality. To do so, we define some statistics and library quality check that we discuss below:

1.3.1 Sequencing and alignment statistics and valid fragments proportion

A high percentage of unmapped reads could indicate a problem in the sequencing run or sample contamination. Normally we expect a percentage of unmapped reads below 10% of total reads. The percentage of chimeric paired reads is an index of long-range ligated fragments. An abnormal value of chimeras in one experiment compare to the others could indicate a problem in the ligation step.

For all our datasets (Hi-C, CHi-C), we got a good proportion of mapped reads (~75%; **Fig 1A**) and most chimeric fragments are valid fragments (**Fig 1B**).

1.3.2 PCR duplicate frequency

A Hi-C library with high quality should have a very small number of PCR duplicates. In fact, the higher the duplication rate, the lower the molecular complexity of the starting library. In general, a Hi-C library with a duplication rate below 20 % of the total library complexity is considered a Hi-C library with good quality (**Fig 1A**).

1.3.3 Inter-chromosomal contacts frequency

Another crucial statistic is the inter-chromosomal contact frequency. A Hi-C library with good quality should have less than 20% of *trans* interactions. A library with high frequency of interchromosomal and with low intrachromosomal contacts suggests that the library contains mostly random ligation products, likely due to the rupture of large fraction of nuclei (**Fig 1C**). Overall, all of our Hi-C and CHi-C datasets passed these quality control steps. Of note, the amount of PCR duplication events was not very different between Hi-C and CHi-C experiments, suggesting that the sequence capture step did not significantly reduce molecular complexity of the library.

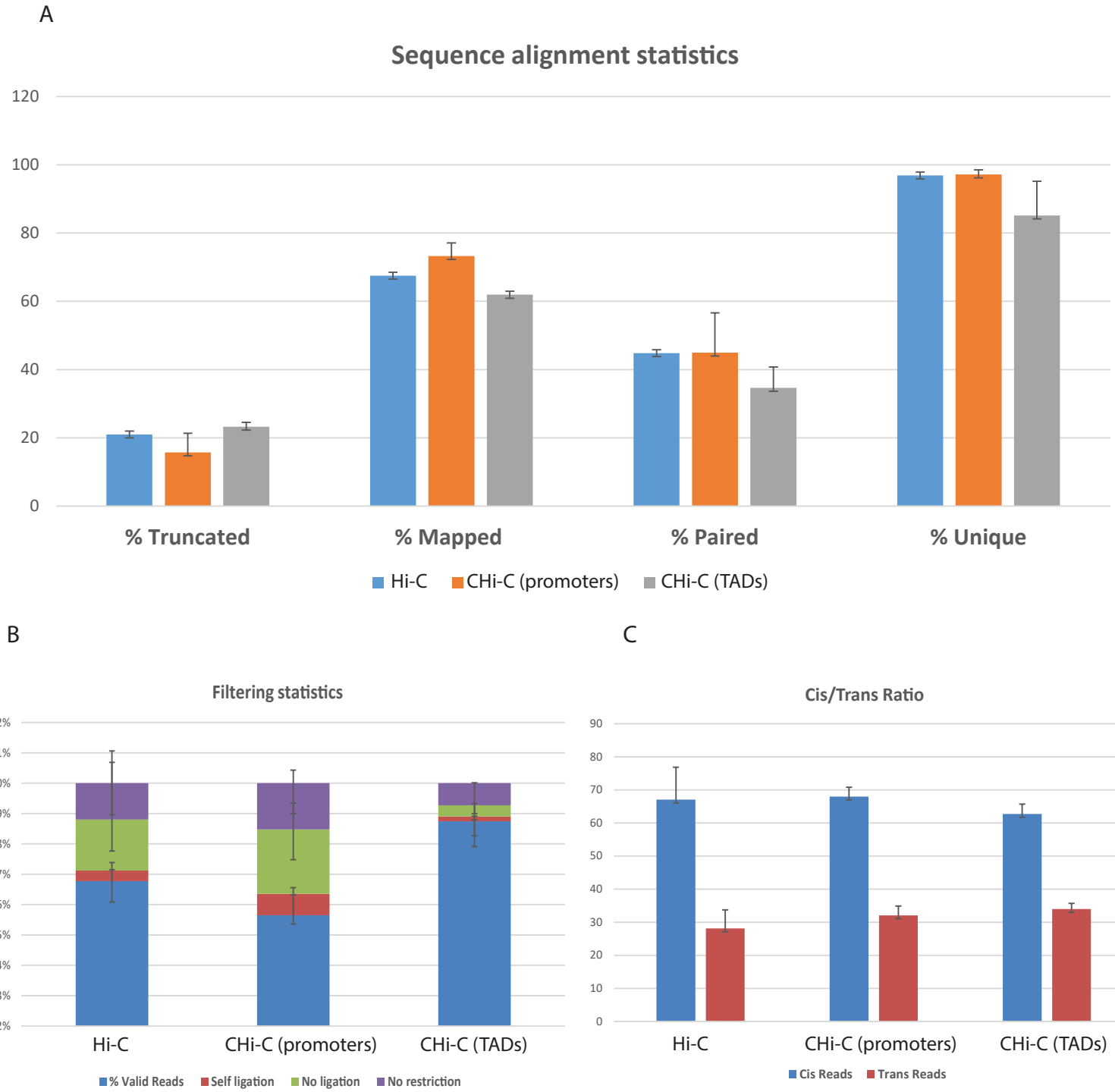


Fig 1: Hi-C library quality controls

A) Sequence alignment statistics for Hi-C (blue), CHi-C promoters (orange) and CHi-C TADs (gray) datasets. For both Hi-C and CHi-C datasets, all sequenced librairies are pooled together (DN3 + DP) .

B) Filtering statistics for Hi-C, CHi-C promoters and CHi-C TADs demonstrating a small fraction of reads are artefacts wherease the highest proportion is for valid reads.

C) Cis/Trans ratios for Hi-C, CHi-C promoters and CHi-C TADs. A small proportion of Trans reads indicates a good quality of sequenced libraries for these different experiments.

2. Evaluation of CHi-C: Promoters and TADs

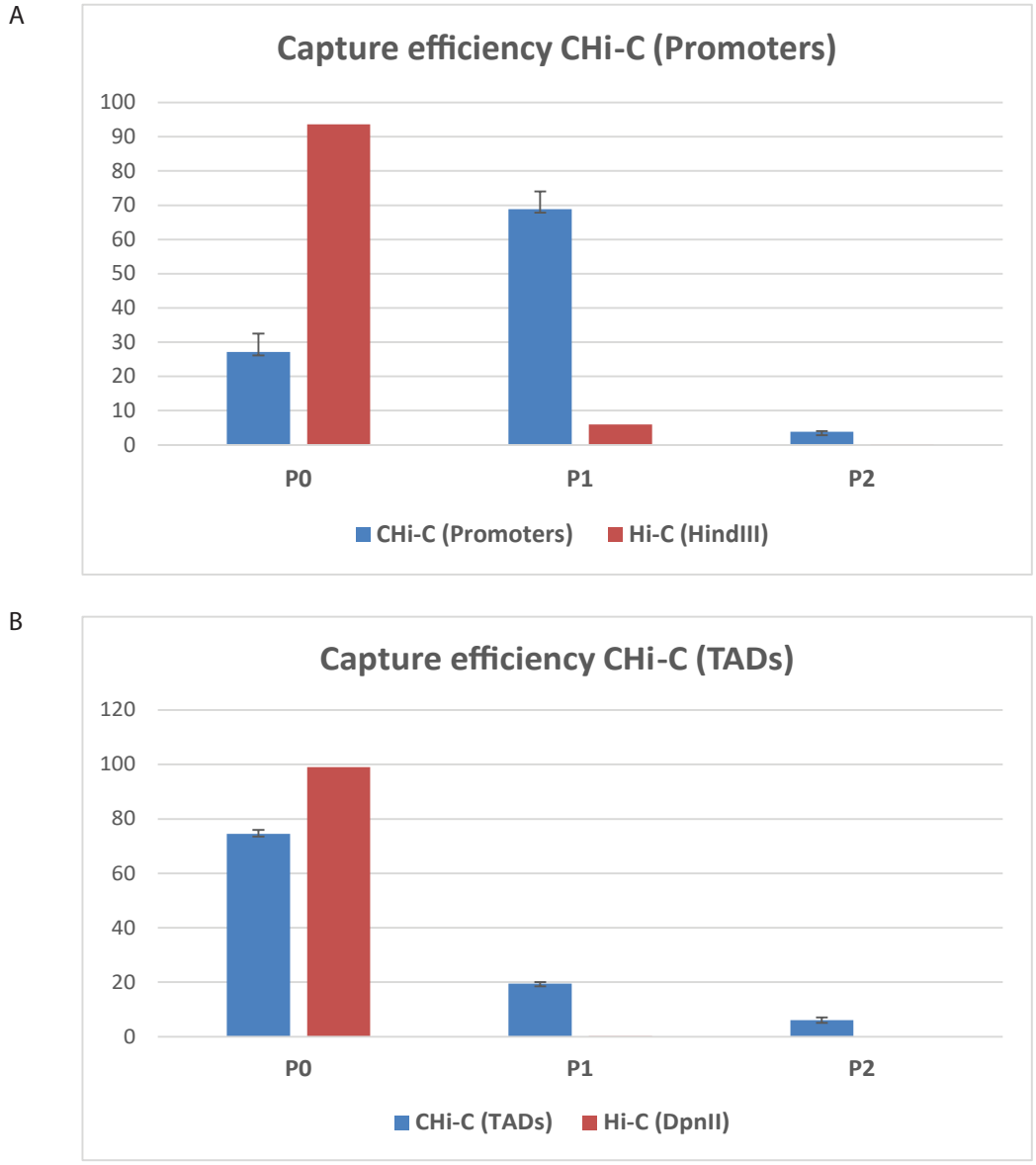
As a further quality control for CHi-C experiments, we calculate the capture efficiency for each experiment. We expect three different populations of chimeras:

- P0: where both chimeric fragments correspond to regions non-targeted by the capture. These are considered failed events of the capture step.
- P1: one of the two fragments is a captured region. These are of the most interest in the promoter CHi-C experiments.
- P2: both chimeric fragments are captured regions. These are of the most interest in the TAD CHi-C experiments.

The capture efficiency is, therefore, calculated as follows: (P1+P2)/ Total number of fragments (**Table 2**).

The P0/1/2 distributions of a CHi-C experiment can then be compared to the "background" distributions arising from the parent Hi-C library (**Fig 2A,B**). In all cases, we observed a high enrichment (~100-fold) of captured products.

We checked also the capture efficiencies of each individual probe within different cell types (DN3, DP, mESCs and FL). The capture efficiency of each probe was calculated as follow: The number of Hi-C reads containing per probe / Total Hi-C reads. Overall, whereas probes do not capture with equivalent efficiency (**Fig 2C**), each bias does not change across cell types, which is expected as the capture is targeting native DNA (**Fig 2D**). Thus CHi-C is much better suited for comparing chromatin topologies across different cell types than ChIA-PET-based methods, where the "capture" biases from immunoprecipitation is not consistent across cell type.



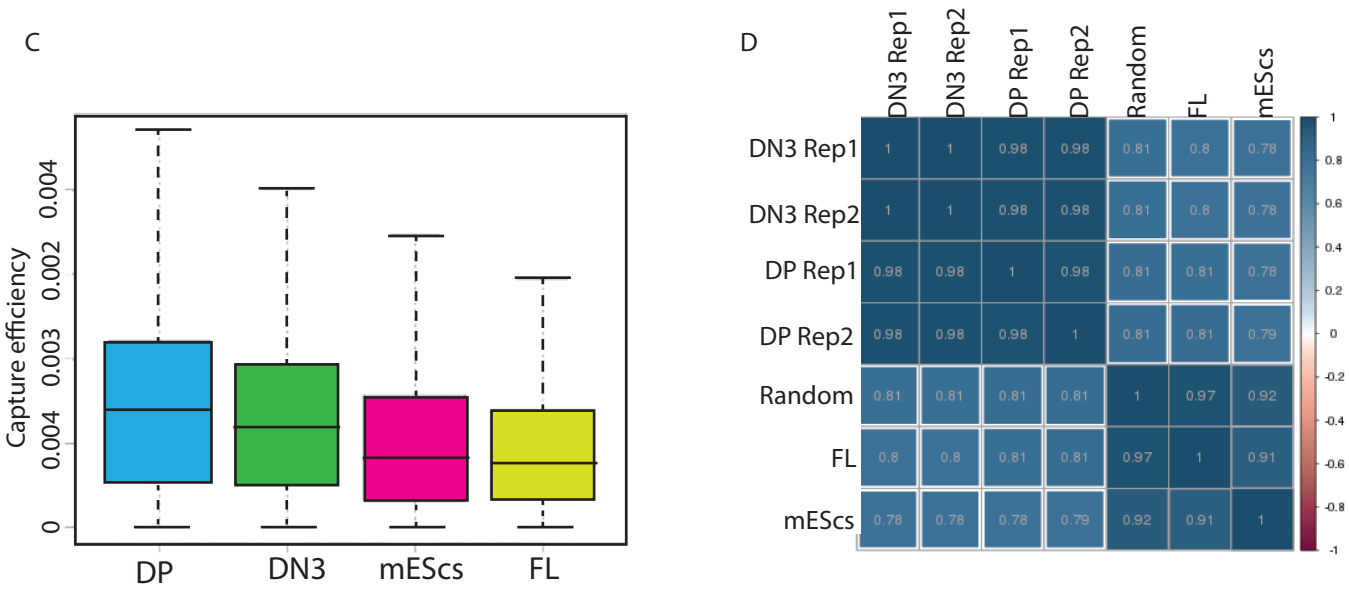


Fig 2: Capture efficiency control

A,B) The Distributions of P0, P1 and P2 CHi-C reads in CHi-C promoters and TADs showing an important enrichment of captured reads (P1+P2) compare to the background reads (P0).
C) Box plot showing individual probe capture efficiency variation for different cell types. The box plot shows an independence of probe capture efficiency from the cell type.
D) Correlation coefficients of individual probe capture efficiencies across different cell types and Native DNA ranomly cut library . The heatmap shows a good correlation between different replicates (lowest is 0.7).

3. Construction of Hi-C contact matrices

In order to have the highest possible coverage of DN3 and DP Hi-C matrices, we combined all datasets for each thymocyte subtype (with HindIII and DpnII enzyme; all biological and technical replicates).

To do so, we converted the mat table of each Hi-C dataset into a text file that describes the alignment of both reads with the ID of the corresponding restriction fragment appended. In this file, each line corresponds to one read pair containing the following fields:

Read name, strand1, chromsome1, position1, fragment1, strand2, chromosome2, position2, fragment2. Then, we use UNIX sort to sort all records in the file with precedence for chromosome, then for fragment, then for strand, and finally for position. This takes on average 30 minutes. The output is then used as input in the juicer pipeline (Durand et al., 2016), more specifically Juicer Tools Pre, in order to construct and normalize contact matrices. Juicer contains three different tools: Juicer Tools Pre, HICCUPS and Arrowhead, which are implemented in Java.

Juicer Tools Pre generates contact matrices at different resolutions (2.5 Mb, 1Mb, 500 Kb, 250 Kb, 100 Kb, 50 Kb, 25 Kb, 10 Kb, and 5 Kb) to give a better view of Hi-C contact matrices. For example, to calculate the contact matrix with 50 Kb resolution, the genome is first linearly divided into 50 Kb bins. Then, for each pair of bins, the number of contacts observed are calculated. Afterwards, Juicer Tools Pre can perform different normalization strategies (see below), on the observed contact Matrix. For our study, we constructed the normalized Hi-C matrix using Juicer Tools Pre with Knight-Ruize normalization. The normalized Hi-C matrix is then stored in a .hic binary file.

4. Hi-C contact matrix normalization

The interaction frequency matrix is very important for making any reliable conclusions on chromatin structure. The generation of robust interaction frequency matrices depends on the Hi-C dataset quality. Ideally, the matrix of observed contacts should be proportional to the true biological contact frequency between two specific loci. However, different studies have demonstrated that Hi-C contact matrices might reflect different biases. So far, three major sources of bias in the Hi-C experiment have been described: the length of restriction fragments, the GC content of sequenced fragments, and the mappability of sequence reads (Cournac, Marie-Nelly, Marbouty, Koszul, & Mozziconacci, 2012; Yaffe & Tanay, 2011). Such effects have been labeled as "one-dimensional biases" that depends on the linear genome:

- Restriction fragment length: The ligation efficiency of restriction fragments depends on the length of restriction fragments, presumably due to different topological constraints on dangling fragment ends of different length to find each other within crosslinked chromatin.
- GC content: A major source of bias in sequencing is the nucleotide composition of the sequence. Reads with very high GC content tends to be more difficult to melt and sequence, and are therefore underrepresented in the final interaction reads.
- Reads mappability: The uniqueness of the fragment ends evidently alters the chances of the fragment passing filters for uniquely mapped reads.

To normalize these biases, different approaches have been used for Hi-C normalization in the literature.

4.1 Vanilla Coverage Normalization

This approach has been used for the first Hi-C matrix generated in the literature. It normalizes the Hi-C matrix coverage by using L_1 norm. Two normalization factors are calculated:

*R_i: the reciprocal of Row sum.

*C_j: the reciprocal of column sum.

Therefore, each normalized entry in the Hi-C contact matrix corresponds to (VC vector): $Observed_{ij}$* R_i* C_j. This approach is easy to implement, highly robust and can be used even with a very sparse data. However, an overcorrection has been noticed with this approach which can be reduced by using the square root of VC vector. Such modification is efficient and very robust compare to the most sophisticated algorithms for Hi-C normalization (Rao et al., 2014b).

4.2 Explicit factor methods

To address the systematic bias cited above, Yaffe and Tanay developed a pipeline that defines a multiplicative probabilistic model based on 3 major biases: GC content, mappability, fragment length. Using maximum likelihood algorithms, they estimate the parameters and then renormalize Hi-C datasets (Yaffe & Tanay, 2011). Similar approaches have been developed after that (for example: HiCNorm (M. Hu et al., 2012)).

4.3 Matrix Balancing

The matrix balancing approach doesn't make any assumptions on which factors are responsible of observed Hi-C biases. The only assumption that is made is that, similarly to Vanilla normalization, observed biases are a one-dimensional multiplicative scalar. The true observed contact matrix is therefore defined as $Observed_{ij}$*C_i*C_j, where C_i and C_j are unknown bias factors. To solve the equation, the row and column sums of the observed matrix are forced to be equal to 1 which means that every locus has the same probability to be observed in the Hi-C matrix. This approach is very well known and used for data analysis. The oldest version of matrix balancing algorithms dates back to the 1930s. A modified version of this algorithm has demonstrated that any square non-negative matrix can be

converted to a stochastic matrix by performing VC normalization on observed entries until convergence is achieved. Recently, this version has been used for Hi-C data normalization (Cournac et al., 2012; Imakaev et al., 2012). With some improvements in algorithm efficiency and speed, Knight and Ruize introduced a new algorithm for matrix balancing based on a different approach for faster convergence.

In general, matrix balancing is an appropriate method for Hi-C data normalization as long as the observed contact matrix is not too sparse.

Since our Hi-C data has a good coverage and was not too sparse, we used for contact matrix normalization the Knight-Ruize matrix balancing approach, implemented in the Juicer pipeline (Durand et al., 2016).

Finally, all these approaches used for Hi-C normalization have high correlation and gives almost the same results with minor differences. All reported features of TADs, peaks or compartments were robust and independent of Hi-C normalization method used (Rao et al., 2014b).

5. TAD CHi-C Normalization

As discussed in the chapter above, CHi-C contains non homogenous populations of fragments that each have their own technical bias. To normalize for possible biases introduced with the capture step, it is necessary to find an approach that takes in account the different biases of Hi-C plus the unknown additional biases of the capture step.

The normalization methods used for Hi-C, specifically matrix balancing approaches, assume that each locus has equal probability or visibility in the Hi-C material. Therefore, this approach is not appropriate for TAD-Capture normalization since it contains heterogeneous material with different visibilities. To overcome this issue, we applied a modified version of iterative correction and eigenvector decomposition (ICE) normalization (Grubert et al., 2015)

exclusively on double captured regions at the specific sub-matrix interrogated. ICE normalization is based on the matrix balancing approach. It corrects collectively for all factors affecting experimental visibility without making any assumptions on possible sources of biases.

The TAD-Capture normalization was accomplished using an R script implementing ICE normalization. The R script depends on two R packages (smoothmest and Matrix) and it normalizes the data iteratively until convergence. For each TAD-capture dataset we have eight matrices of P2 results that correspond to 8 loci targeted by the capture. Each matrix was normalized using a maximum of 20 iterations, since we got very fast convergence (around iteration 14) which took roughly 30 seconds. The output file is in the ibed format, where each line corresponds to a pair of bin interactions. It contains the following fields: chr_bin1, start_bin1, end_bin1, chr_bin2, start_bin2, end_bin2, Observed value, Normalized value.

Table 1: Total sequencing reads

	Hi-C	**Promoter-capture**	**TAD-capture**
DN3	HindIII: 509,021,728 DpnII: 1,022,206,750	2,162,141,732	927,642,574
DP	HindIII: 575,247,926 DpnII:1,018,316,264	2,047,176,034	1,032,414,172
mESc	1,379,277,446		1,370,739,900

Table 2: Captured reads after filtering of CHi-C (Promoters)

	P1	**P2**	**Total**
DN3 Rep1	165,950,037	9,443,317	175,393,354
DN3 Rep2	150,733,333	8,091,686	158,825,019
DP Rep1	140,127,943	7,830,171	174,958,114
DP Rep2	97,878,747	5,669,171	103,547,918

PromoMaxima: a pipeline for detection and visualization of *cis*-DNA looping in Capture Hi-C

Yousra Ben Zouari[1-4], Dominique Kobi[1-4] and Tom Sexton[1-4]

[1] Institute of Genetics and Molecular and Cellular Biology (IGBMC), Illkirch, France

[2] CNRS UMR7104, Illkirch, France

[3] INSERM U1258, Illkirch, France

[4] University of Strasbourg, Illkirch, France

Abstract:

Capture Hi-C (CHi-C) is a new technique developed for assessing genome organization. It is based on chromosome conformation capture techniques, involving a capture of regions of interest such as gene promoters. CHi-C data analysis is a challenging task since neither existing Hi-C-like nor 4C-like analyses are suitable, making different assumptions about the technical biases presented. We describe a new method for CHi-C analysis, PromoMaxima, which shows more stringency and robustness compare to previously developed CHi-C analysis tools. It uses local maxima combined with a background model to detect DNA looping interactions in Capture Hi-C data, and flexibly integrates information from biological replicates. The tool is also presented with a ready-to-use browser, allowing visualization of CHi-C data alongside linear epigenomic profiles, such as ChIP-seq data. The PromoMaxima R scripts will soon be available on Github.

Key words:

Promoter-enhancer interactions, Chromatin loops, Capture Hi-C

Background

The advent of the chromosome conformation capture (3C) technology(Dekker et al., 2002) allowed higher-order chromosome folding to be inferred by identifying spatial proximity between distal genomic sequences, leading to a comprehensive insights of genome topology. As sequencing throughput has increased, it has become feasible to globally assess all chromatin interactions within a population (4C: “one-to-all”, 5C: “many-to-many”, Hi-C: “all-to-all” methods) simply by sequencing 3C ligation products(Dostie et al., 2006; Fullwood et al., 2009; Hughes et al., 2014; Lieberman-Aiden et al., 2009a; Mifsud et al., 2015; Simonis et al., 2006). In fact, Hi-C interaction maps can give insight into chromosome folding at different scales, depending on the sequencing depth (and hence resolution) of the study.

However, the strength of Hi-C in assessing all possible chromatin interactions is also one of its major disadvantages: the numbers of possible ligation products that can be detected is much greater than current sequencing output. Recently, several groups have coupled Hi-C (or another 3C derivative) to sequence capture with pools of oligonucleotides complementary to thousands of restriction fragment ends(Dryden et al., 2014; Jäger et al., 2015; Mifsud et al., 2015; Schoenfelder, Furlan-Magaril, et al., 2015a; Schoenfelder, Sugar, et al., 2015). Such "CHi-C" methods allow interactomes for large subsets of the genome, such as all promoters or DNase hypersensitive sites, to be simultaneously mapped at higher resolution. Although highly informative, CHi-C datasets have specific properties that set them apart from other 3C-like techniques, and so require specialized analytical tools. The majority of CHi-C strategies involve large numbers (thousands) of spatially dispersed baits which lead to an asymmetry of CHi-C contact matrices. In addition, individual baits have variable capture efficiencies which introduce additional technical biases. Depending on the bait design, CHi-C datasets will be more or less populated with ligation products between two bait fragments, as well as between bait and non-bait, which may complicate bias assessment even further.

As for all genome-wide datasets, the challenges for CHi-C analysis are in the appropriate definition of an expected background level, from which "significant" signal can be resolved, and correct normalization to non-biological biases. Up to now, two methods have been used for CHi-C analysis: CHiCAGO (Cairns et al., 2016) and GOTHiC (Mifsud et al., 2017). GOTHiC, actually developed for interaction calling in Hi-C, employs a very simplistic binomial test coupled with multiple testing correction to search for over- represented interactions, but makes no consideration for known features of Hi-C data, such as the heavy dependence of "background" interactions on genomic distance, let alone aspects of CHi-C such as capture bias. CHiCAGO uses a statistical background model to account for different biases in promoter-CHi-C data, combining three factors to define the expected

background interaction level: genomic distance, bait capture efficiency, and technical biases present in Hi-C and sequencing approaches. These parameters are fit to the data to define an expected interaction strength for each individual restriction fragment, based on a combined negative binomial and Poisson variable. However, the treatment of each single fragment as an independent variable creates problems when accounting for biological replicates, since despite its improved coverage compared to Hi-C, current depths of CHi-C datasets still vastly sub-sample the possible space of ligation products. As a result, many reproducible chromatin loops observed at the resolution of larger bins of pooled restriction fragments are lost when scoring individual restriction fragments (**Fig S1**). CHiCAGO utilizes the same geometric mean approach as DESeq2(Anders, 2014) to allow weighting for different read depths of different replicates, but this is insufficient to completely counter the problem. Further, chromatin interactions comprising contiguous fragments of increased signal, centered on an interaction peak, are less likely to result from technical artefacts than isolated "spikes" of signal. We tried to overcome these existing limitations of CHi-C analysis methods, and developed PromoMaxima, which we applied to published mouse embryonic stem (ES) cell promoter CHi-C data (Schoenfelder, Furlan-Magaril, et al., 2015b) and benchmarked against GOTHiC and CHiCAGO.

Results

Methodological foundation of PromoMaxima

In '3C' approaches, genomic distance has an important impact on the expected frequency of interactions. Generally, the frequency of interactions decays on a power law scaling as the genomic distance between fragments increases, consistent with many polymer physics models(Lieberman-Aiden et al., 2009b). DNA loops correspond to a peak or a higher signal (hills) of interactions compared to the expected level of neighbor fragments (valleys) on either

side; such features were used to detect loops in the first 3C studies(Palstra et al., 2003). To detect these hills, we use a non-parametric approach used for detection of signal peaks, namely the local maxima, without making any pre-assumptions or preconceived model of the data (**Fig 1**).

Specifically, treating each bait independently and removing bait-to-bait interactions, we obtain a "virtual 4C" profile of read counts relative to the genomic position of the non-bait fragment, and perform loess smoothing on this profile. The fragments with the maximum signal are identified within sliding windows of a given number of fragments, and local maxima are defined as regions where the smoothed signal equals this value. In this approach, two parameters need to be controlled: the span of the loess smoothing (*s*), and the window size (*w*) for the local maximum computation. Over-smoothing or using too large a window size causes maxima to not be called, and under-smoothing or small window sizes call many local spikes as spurious interactions. ROC (receiver operating characteristic) analysis found smaller *s* and larger *w* to be optimal (**Fig S2**). However, very small local maxima, which are very distant from the bait and so have a negligible background signal, are still called as "interactions" (**Fig 1**), which needed to be filtered by a better estimation of the background model.

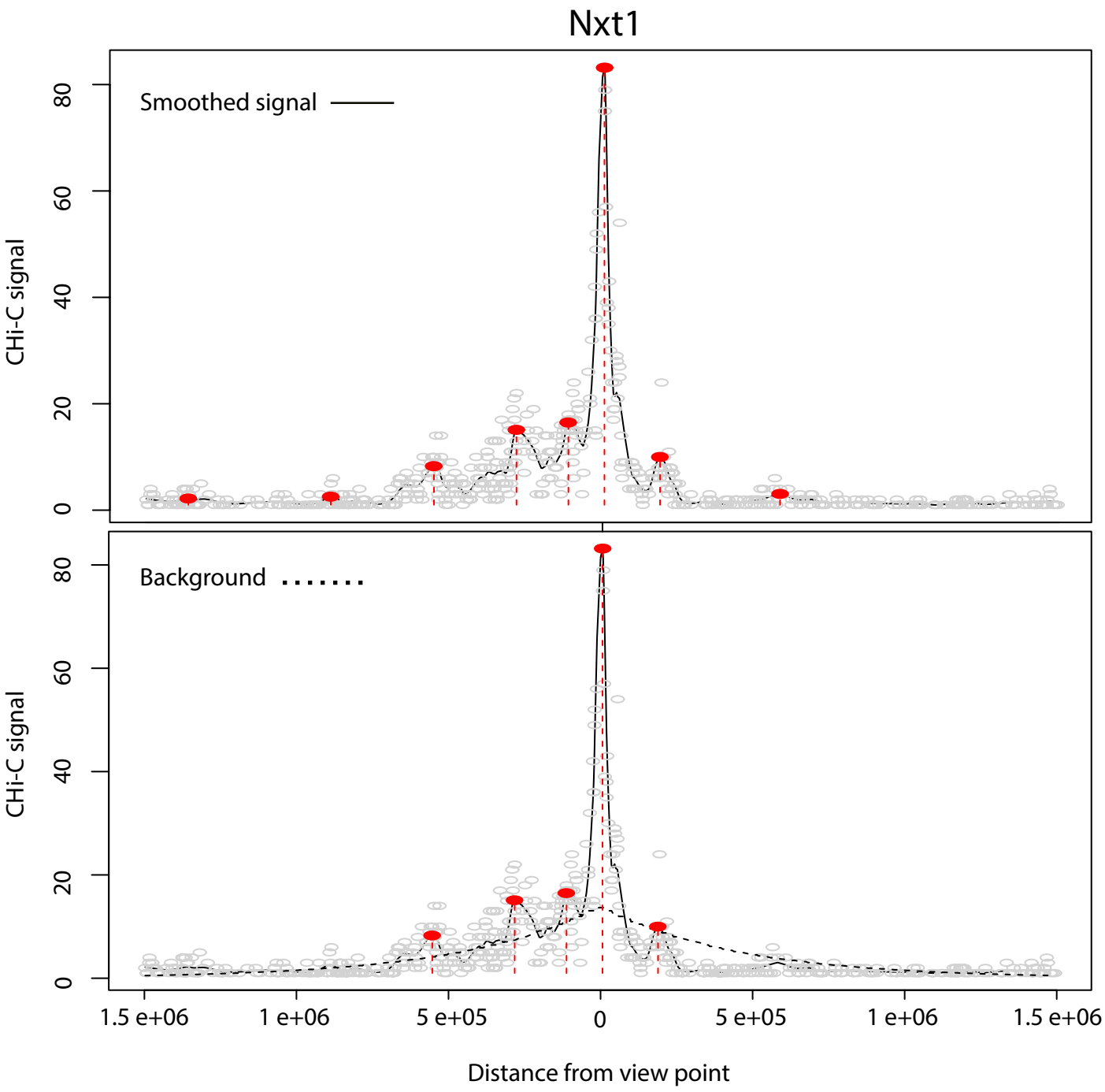


Fig 1: Called interactions by PromoMaxima for Nxt1 gene in mESCs.
Top: local maxima (red spots) are called from loess smoothed profiles (black line); white circles represent raw reads.
Bottom: Background distance model (dotted line) is applied to filter out spuriously called interactions.

Estimation of the background level

According to previous work on CHi-C data(Cairns et al., 2016), the background interaction level at short genomic distances (up to ~1.5 Mb) is largely dominated by genomic separation (proposed to be caused by Brownian collisions of the chromosome fiber). In CHiCAGO, a cubic-fitted distance function was derived from the geometric means of read counts for binned genomic separations, and was then scaled with capture bias estimates in the final derived background distribution(Cairns et al., 2016). Inspired by this, we derived similar but *bait-specific* genomic distance functions, fit to each virtual 4C profile. Instead of a cubic fit, we applied a fit to a negative binomial distribution, to account for the known overdispersion of sequencing data(Anders, 2014). We found that filtering local maxima for those whose signal exceed the background level easily removed likely false negatives (**Fig 1**).

Accounting for biological replicates

Although CHi-C improves on the resolution afforded by conventional Hi-C, it remains an under-sampled method, explaining the poor reproducibility of called interactions at restriction fragment level between biological replicates (**Fig S1**). Although taking the intersection of called interactions from each replicate will give the highest-confidence chromatin loops, the false negative rate appears to be very high from this approach, even if PromoMaxima appears to perform better than CHiCAGO. To add more flexibility, PromoMaxima allows a window size between reported peaks in biological replicates to be defined by the user (*w*: default is 0 kb). Background model-filtered local maxima are computed for each biological replicate, and high-confidence interactions are called as those that have a called interaction in both replicates within *w* bp of each other. Empirically, most called peaks within biological duplicates in the ES promoter CHi-C data were contained within 30 kb (~7.5 *Hind*III restriction fragments) of each other (**Fig S3**). We used this window size for subsequent analyses, but note that the majority of replicate-consistent interactions are much closer.

Benchmarking of PromoMaxima

Having defined the optimal parameters for PromoMaxima, we performed it on a published mouse ES promoter CHi-C dataset(Schoenfelder, Furlan-Magaril, et al., 2015b), and compared our results with those of CHiCAGO and GOTHiC (**Table 1**). On visual inspection, PromoMaxima successfully identified clear promoter interactions, which had also been validated by 4C, and seemed to call fewer spurious ones than the other two methods (**Fig 2**). Indeed, PromoMaxima identified fewer promoter-centered interactions (24,488) than CHiCAGO (94,148) or GOTHiC (548,551). Notably, the vast majority of PromoMaxima-called interactions were recapitulated in the other two methods (75% by CHiCAGO; 83% by GOTHiC; **Fig 3A**), suggesting that PromoMaxima is the most stringent method but also calls the highest-confidence interactions, and likely has a lower false positive rate. In support of this, the PromoMaxima-called interactions also had significantly higher interaction score metrics as called by the other two techniques (observed/expected ratios for GOTHiC; weighted CHiCAGO scores) than interactions called by either of the other two techniques but not by PromoMaxima (**Fig 3B**; $P < 2x10^{-16}$, Wilcoxon rank sum test in both cases).

One of the major perceived applications of CHi-C is to assign target genes to candidate *cis*-regulatory elements, particularly enhancers, by virtue of the specific interactions they make with promoters. Genomic studies revealed that enhancers share hallmark chromatin features: monomethylation of histone H3 lysine-4 (H3K4me1), DNase-hypersensitivity, acetylation of histone H3 lysine-27 (H3K27ac) and/or p300 co-activator occupancy(Heintzman et al., 2009). However, despite epigenomic predictions of enhancers in numerous cell types, unambiguous identification of their target genes has proved more elusive, since they can control multiple genes, and may skip one or several promoters to act over large distances(Sanyal et al., 2012). Promoter CHi-C studies have indeed shown a general enrichment in interacting regions bearing enhancer chromatin signatures(Hughes et

al., 2014; Sahlén et al., 2015; Schoenfelder, Furlan-Magaril, et al., 2015), as well as for regions bound by CTCF, a known factor implicated in chromatin loops(Phillips & Corces, 2009). We reasoned that an interaction calling method that found the greatest proportion of putative enhancers and/or CTCF sites within a promoter CHi-C dataset was most likely to have the best true positive detection rate. Based on this, PromoMaxima compares favorably to the other two methods. It has a higher enrichment for interacting regions containing CTCF, H3K27ac and H3K4me1 (**Fig 3C**), with a ~2-fold improvement over CHiCAGO and ~5-fold improvement over GOTHiC. Conversely, we assessed which of the 19,201 candidate mouse ES enhancers (based on chromatin signatures(Shen et al., 2012)) could be assigned to target promoters by the different methods (**Table S1**). As expected, the proportion of assigned enhancers scaled with the numbers of total called interactions (68% for GOTHiC; 25% for CHiCAGO; 21% for PromoMaxima). However, candidate enhancers comprised a much higher proportion of the PromoMaxima-called interaction set than for the other two methods (~3-fold higher than CHiCAGO; ~6-fold higher than GOTHiC; **Table S3**), in line with the relative enrichments for individual regulatory marks. We note that promoter interactions with non-enhancer/CTCF-bound elements may certainly be frequent and functionally significant, albeit poorly characterized so far. Indeed, all three methods call many interactions of this category. However, the greater enrichment of PromoMaxima-called interactions for promoter-enhancer loops that have been so well described in the literature, coupled with their overall higher interaction score metrics as called by other methods, suggests that PromoMaxima is the most stringent interaction calling method, but also reliably identifies the interactions most likely to be functionally relevant.

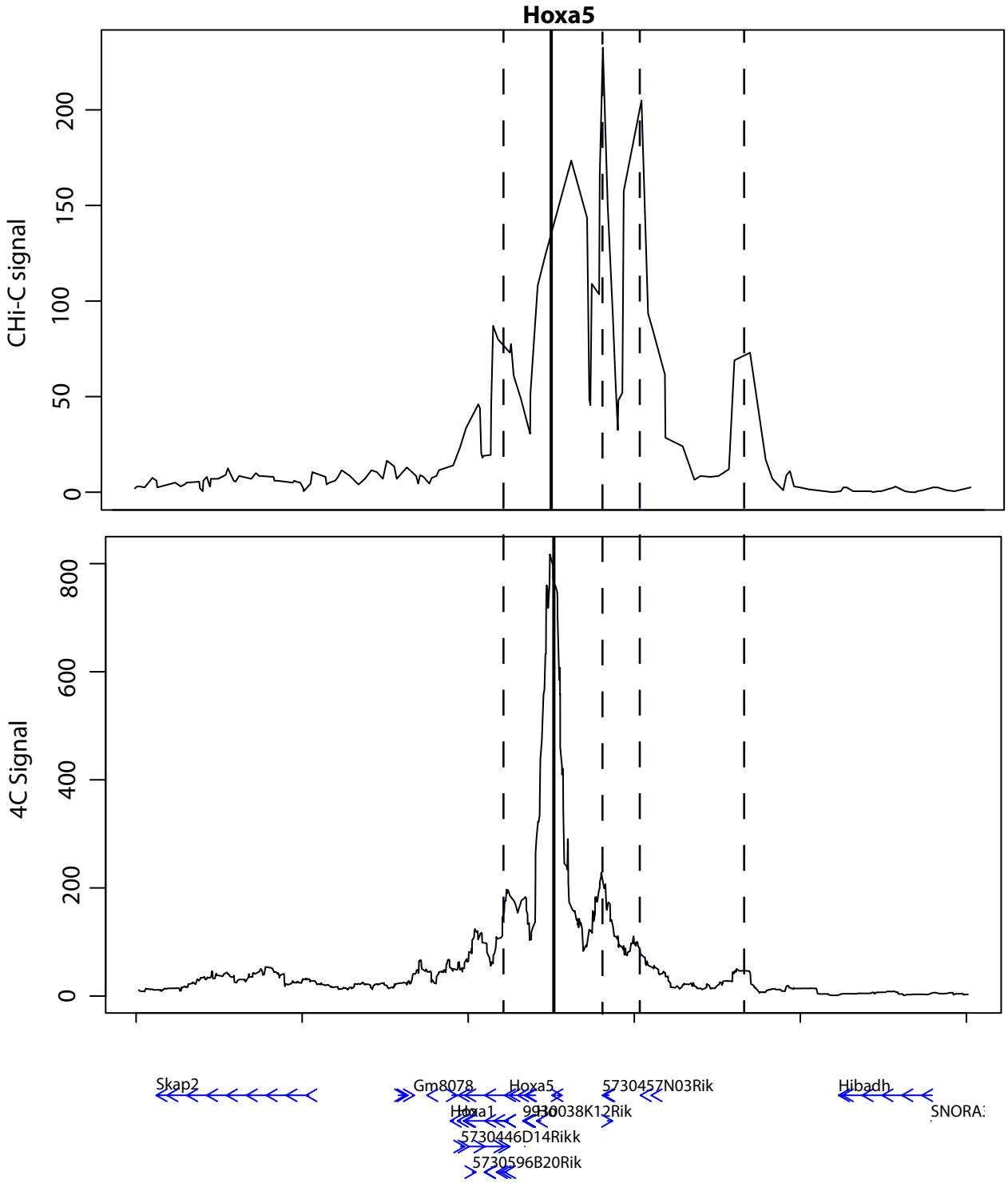


Fig 2: Example of DNA loops, called with PromoMaxima for the gene Hoxa5, and validated by 4C in mESCs.

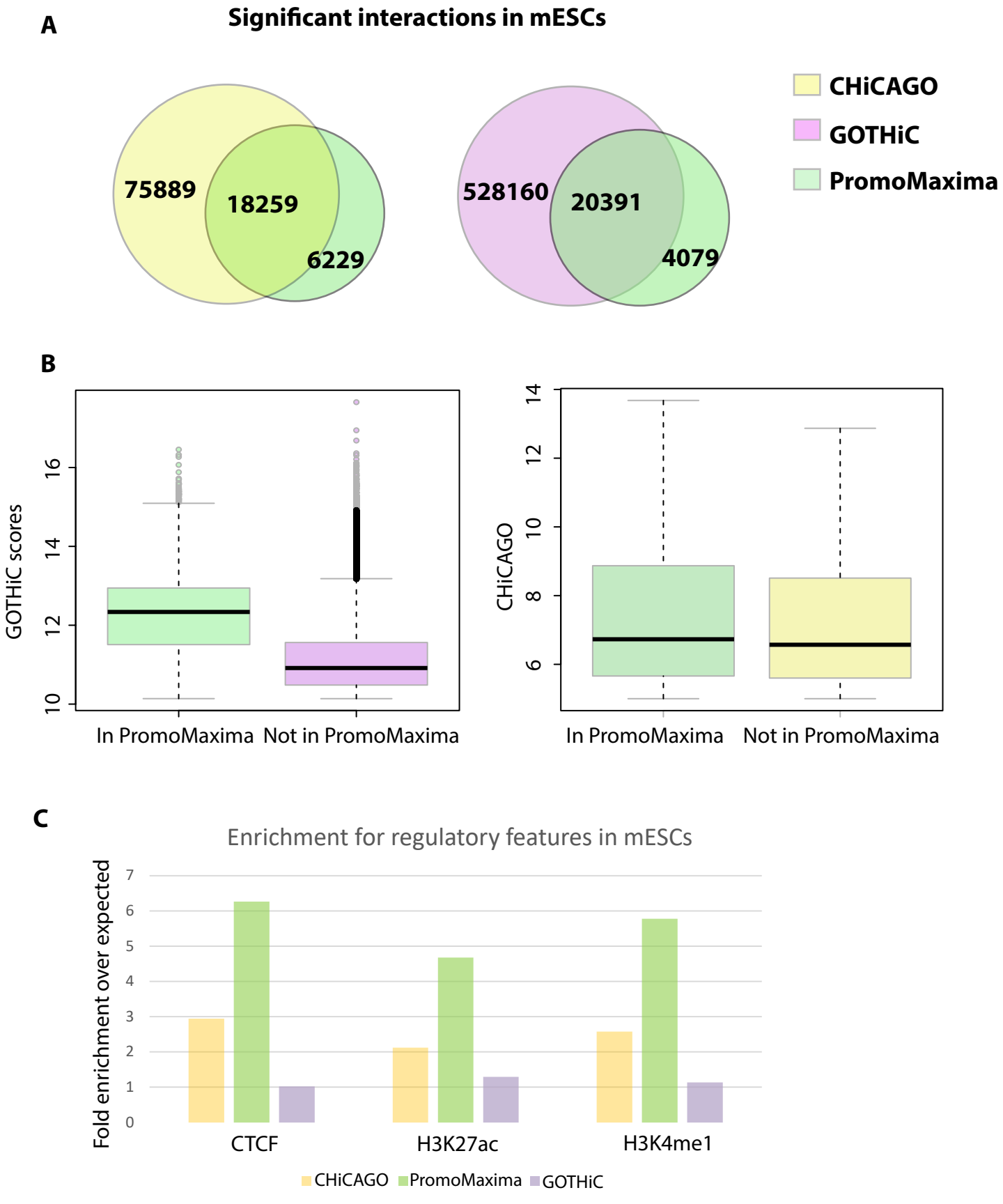


Fig 3: Benchmark of PromoMaxima, GOTHIC and ChiCAGO.
A. Venn diagram of GOTHiC, CHiCAGO and PromoMaxima called interactions.
B. Boxplots comparing GOTHiC and CHiCAGO scores for interactions called or not by PromoMaxima.
C. Enrichment of epigenetic mark peaks in CHiCAGO, GOTHiC and PromoMaxima called interactions.

PromoMaxima Browser

For ease of visualization of all promoter CHi-C results, as well as the interactions called by PromoMaxima, we also present an R-based browser. Unlike the WashU browser(Zhou et al., 2013), which displays all interactions simultaneously and can be quite difficult to interpret visually, the PromoMaxima browser displays bait-specific virtual 4C profiles, with the bait (target gene) and display window (as precise genomic coordinates, or a window size flanking the bait) specified by the user via a graphical window (**Fig 4**). The entire CHi-C dataset(s) is called up once in the random-access memory, and is then used to generate the user-specified plots rapidly. The loess quantile normalized virtual 4C profile is plotted, and the browser provides the option to highlight multiple interactions lists (replicates, biological conditions, different calling methods). Linear epigenomic datasets, such as ChIP-seq tracks, can also be uploaded as bigWig files for direct comparison with the CHi-C results. Altogether, PromoMaxima can be used to fairly compare the CHi-C profile of different conditions or cell types.

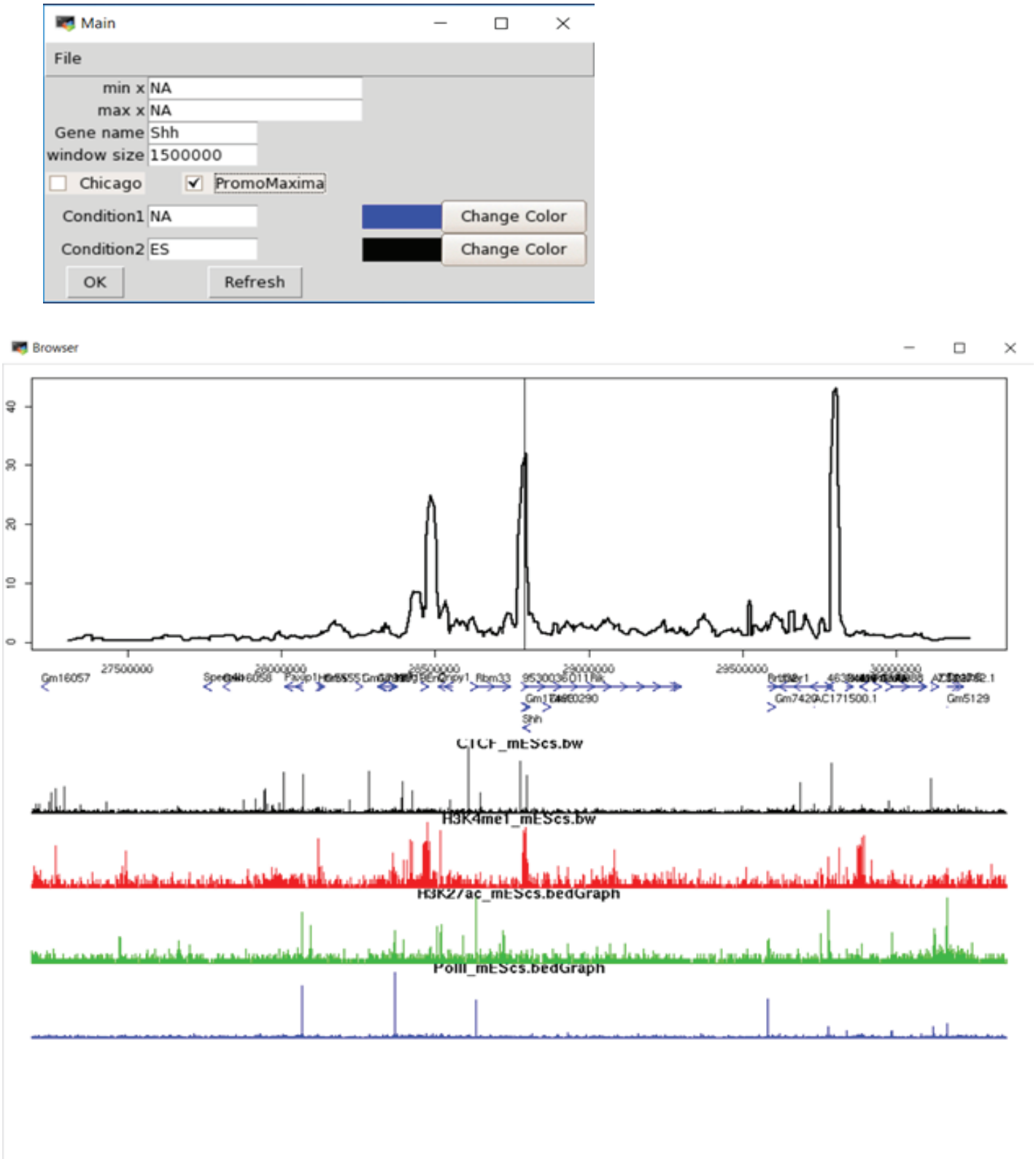


Fig 4: PromoMaxima Browser

A screen shot of PromoMaxima browser showing *Shh* CHi-C profile in mouse ES, with ChIP-seq profiles underneath. We distinguish two important DNA loops that concomitant with H3K4me1 peaks, indicating the presence of potential enhancers.

Discussion

We present the PromoMaxima tool for Capture Hi-C analysis and visualization. PromoMaxima demonstrates an efficient detection of interactions enriched for regulatory chromatin features. Compared to existing tools, PromoMaxima showed more stringency in called interactions and better enrichment of regulatory features. So far, PromoMaxima has restricted analyses to close intrachromosomal (within 1.5 Mb of the bait) interactions between bait and non-bait sequences. Longer-range and interchromosomal interactions are generally much weaker(Lieberman-Aiden et al., 2009b), and more difficult to robustly detect. Importantly, the background is very close to zero at current sub-saturation sequencing depths, so even noise/weak signals appear as local maxima; PromoMaxima may thus not be an appropriate method for investigating these types of interactions. On the contrary, close intrachromosomal bait-to-bait interactions (i.e. between promoters) could conceivably be assessed by PromoMaxima, but a much greater analysis of the interplay between bait capture biases will be required to assess its robustness, and is omitted from the current pipeline. Overall, PromoMaxima is a useful, stringent yet robust, tool for calling promoter (or other sparsely dispersed CHi-C bait) interactions from CHi-C data, provided with a user-friendly graphical browser for visualization of the results.

Methods

PromoMaxima Pre Processing

The input of PromoMaxima is an ibed file format which contains these columns: bait ID, Chr Bait, Start Bait, End Bait, OE ID, Chr OE, Start OE, End OE, Number of reads.

The input contains the coordinates of interacting regions with their corresponding counts for each condition. It is important to keep the order and the headers specified in the README file. A settings file is coupled with the PromoMaxima pipeline in which the user defines the different biological conditions (the labels of added columns for each condition).

PromoMaxima is an R script designed to be integrated later as Bioconductor Package. It depends on R version >= R 3.0.2 and the pre installation of these packages: Rsamtools, GenomicRanges, limma, caTools, data.table, base, zoo, RcppRoll and psych. An error message will be thrown if any of these packages are missing.

PromoMaxima is called by […], with the following user-provided arguments possible:

Usage:

./PromoMaxima.R [options]

Options:

-i/--input [default:input] #The ibed inupt file

-o/--output [default:./ouput] #The output folder

-d/--distance [default: 0 bp] #The distance between biological replicates

-s/--settings [default: path of setting file] #The setting file path

-w/--window [default: 50] # The window of loess smooth

-sp/--span [default: 0.05] # The span if loess smooth

PromoMaxima then performs, for each bait:

- Extraction of intrachromosomal and bait to non-bait for background level estimation interactions within 1.5 Mb.
- Loess smoothing with two parameters to define (window and span of the loess smooth). We used w=50 and span 0.05
- Local maximum calling.
- Background model estimation; see below for details.
- Intersection of called peaks within a user-defined window (distance parameter); default is zero, and for this analysis we used 30000.
- The output is a file in the ibed format of identified interactions for each biological condition.

Estimation of the background level

To estimate the background level, first we binned the genomic distance from bait up to 1.5 Mb into bins of 20 kb (approximately 5 HindIII restriction fragments). For a given bait, we first calculate the average count over all of the other ends excluding baits (promoter-promoter interactions) whose distance from bait falls in a given bin (removing all zeros to eliminate numerical instabilities). The function distance then is estimated base on the geometric mean of all bins at that distance using negative binomial regression (glm.nb). Predicted values are then used to account for the background level. Detected peaks with local maxima are then reported if they exceed the background level.

ROC analysis

Using different values for window and span parameters, we identified peaks in different bait pools. Then, for each interacting restriction fragment a binary value is assigned to indicate if it is a peak or not. Using the R package ROCR, we plotted the different ROC curves for each set of parameters. We set the label as the presence/absence of peak and the read counts as corresponding values.

Benchmarking PromoMaxima with other tools

To benchmark PromoMaxima with CHiCAGO and GOTHIC we downloaded the list of called interactions by these tools from GEO (**Table S1**).

To determine the Jaccard index between biological replicates, we run CHiCAGO on CHi-C data from GEO (2 biological replicates; **Table S1**).

4C datasets

The 4C data of Hoxa5 on mES cells are produced in our lab, following an established protocol (Noordermeer et al., 2014). To process the data of 4C we used a customized pipeline, adapted from those previously published (de Wit et al., 2015b). The 4C profile is then plotted by applying a running mean on 21 restriction fragments.

Epigenetic marks and ChIP seq analysis

ChIP-seq data were downloaded from GEO database for these histone marks: H3K4me1, H3K27ac and CTCF (**Table S1**). We aligned ChIP-seq data to mm9 genome using bowtie2, then called peaks using Erange V4.0. In Erange, mapped reads in the SAM file are first transformed into native Erange reads stored as an .rds file. Then, peaks are identified with the peaks finder tools in Erange with –nodirectionality and –notrim. Erange returns a per-peaks p-value. By default, this is calculated using a Poisson distribution of peak reads per million base pair (RPM) for each chromosome (FDR <= 0.05). Enrichment of each epigenetic feature

within an interaction set was computed by dividing the proportion of interactions overlapping with a feature peak within the interaction set by the proportion of all restriction fragments which overlap with a feature peak.

PromoMaxima browser

PromoMaxima browser is a user friendly tool for visualization of CHi-C profiles. It depends on these libraries: tcltk2, tkrplot, limma, caTools, data.table, base, zoo, rtracklayer, AnnotationHub, Rsamtools, gplots and R version >= 3.2.

The GUI features present in the browser for the user to play with are:

- Gene name
- Window size in base pair to be plotted (symmetric from the view point)
- The min/max genomic coordinates of the plot in case the user is interested in a specific genomic region
- Upload the list of called interactions by PromoMaxima/CHiCAGO to visualize
- Different biological conditions/cell types with the corresponding colors
- Upload the epigenetic profiles (bigwig files)
- Save a screenshot from the browser(eps format)

Supplementary data

Table S1: GEO datasets

	GEO data
CHi-C data of mouse ES	**E-MTAB-2414 (ArrayExpress)**
CTCF	**GSM723015**
H3K4me1	**GSM1359829**
H3K27ac	**GSM851278**
CHiCAGO interactions of ES	**GSE81503**
GOTHiC interactions of ES	**E-MTAB-2414 (ArrayExpress)**

Table S2: Comparison of CHiCAGO, GOTHIC and PromoMaxima detected interactions in mESCs

	CHiCAGO	**GOTHiC**	**PromoMaxima**
Number of captured baits	**22,459**	**22,459**	**22,459**
Number of significant interactions	**94,148**	**548551**	**24,488**
Mean number of significant interactions per bait	**4.19**	**29.4777**	**1.7**
Median distance of cis-chromosomal interactions	**155,200 bp**	**34,776 bp**	**138,077 bp**

Table S3: Overlap of called interactions by different tools and called Enhancers based on their epigenetic features

	Enhancers mESCs (19201)
PromoMaxima	**21% (3993)**
CHiCAGO	**25% (4777)**
GOTHiC	**68% (13219)**

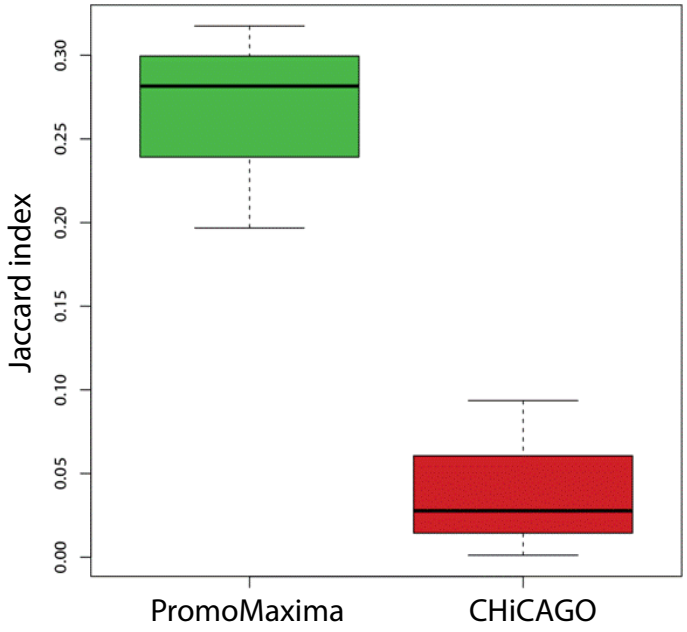


Fig S1: Jaccard index of called interactions maintained in biological replicates, using PromoMaxima and CHiCAGO on three different promoter CHi-C experiments.

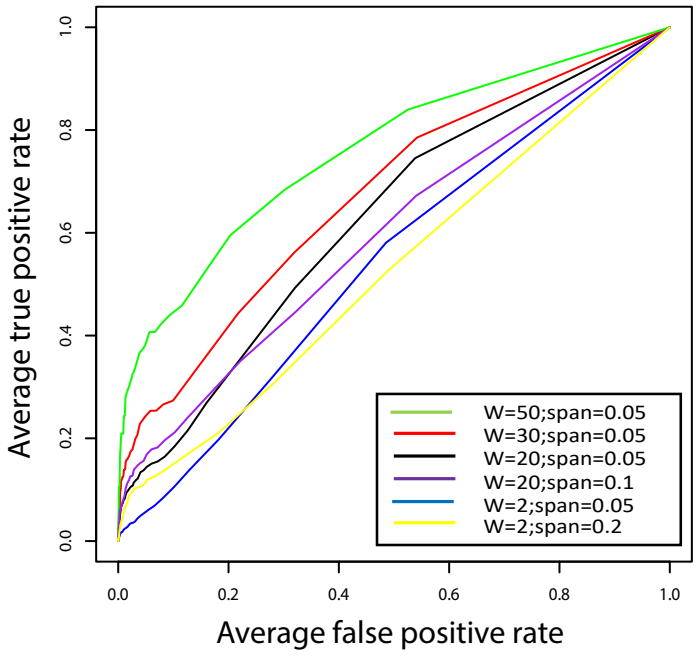


Fig S2: ROC curve of window size and span parameters for local maxima calling in PromoMaxima.

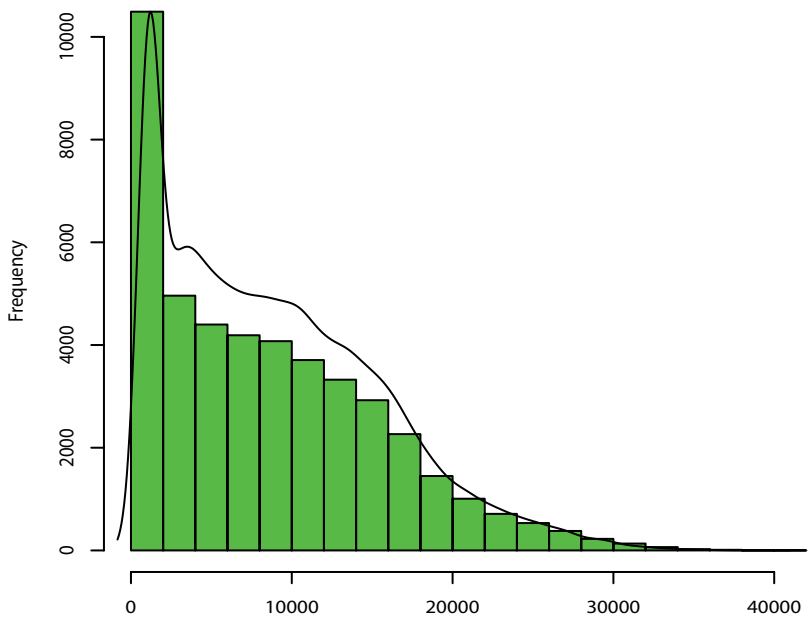


Fig S3: Distance between interaction peaks in biological replicates of CHi-C data from mESCs (bp)

-

Developmentally dynamic gene promoter interactions in transcriptional activation and repression

Anne M Molitor[1-4,6], Yousra Ben Zouari[1-4,6], Dominique Kobi[1-4], Manon Maroquenne[1-4], Audrey Mossler[1-4], Sanjay Chahar[1-4], Nezih Karasu[1-4], Stefan Schoenfelder[5], & Tom Sexton[1-4]*

[1] Institute of Genetics and Molecular and Cellular Biology (IGBMC), Illkirch, France

[2] CNRS UMR7104, Illkirch, France

[3] INSERM U1258, Illkirch, France

[4] University of Strasbourg, Illkirch, France

[5] Babraham Institute, Cambridge, UK

[6] These authors contributed equally to this work.

* Correspondence should be addressed to T.S. (sexton@igbmc.fr).

Preface

This work was a large collaborative effort within the group. I performed all of the computational analysis of the CHi-C data and comparison with all other genome-wide datasets. I did not perform any of the experimental follow-up work, although the candidate genes/regions were chosen from the results of my analysis, and in consultation between myself and other members of the group.

Abstract and Introduction

Tight regulation of metazoan transcription underpins developmental fate decisions. Such control is not solely conferred by gene promoters, but involves combinatorial input by distal *cis*-regulatory elements. Activating enhancers are by far the best understood, due to their detailed epigenomic characterization (Roadmap, 2015), and extensive discovery of chromatin looping interactions with their target gene promoters (Palstra et al., 2003; Sanyal et al., 2012; de Wit et al., 2013; Hughes et al., 2014; Sahlen et al., 2015; Schoenfelder et al., 2015a; Mifsud et al., 2015). However, it is still unclear how chromatin topology is linked to transcriptional activation, due to conflicting reports of developmentally constitutive (Ghavi-Helm et al., 2014; Hakim et al., 2011; Jin et al., 2013) and dynamic (Palstra et al., 2003; Javierre et al., 2016; Rubin et al., 2017; Siersbaek et al., 2017) enhancer-promoter interactions. In contrast, although distal gene silencer elements were described decades ago (Kadesch et al., 1986; Sawada et al., 1994), their genome-wide prevalence and DNA sequence/chromatin signatures remain largely unassessed. Further, in the few cases where chromatin topology around described silencers has been addressed, they appear to sequester enhancers away from target promoters than to directly interact with the promoters themselves(Jiang & Peterlin, 2008; Jing et al., 2008). We systematically interrogated promoter interactions during mouse thymocyte development, and uncovered an extensive and complex cell type-specific spatial network of both maintained and dynamic interactions with functional enhancers *and* silencers. As may be expected, regions interacting with active genes were enriched in enhancer marks, such as H3K4me1 and H3K27ac. However, no epigenetic signature appeared to distinguish enhancers participating in stable or dynamic loops; “poised” and active enhancers were as likely to form contacts with target genes. Functional silencers were devoid of enhancer signatures, but were also not highly enriched in the classical hallmarks of repressed chromatin, such as H3K9me3 or Polycomb-mediated H3K27me3. A

significant number of silencers comprised LINE repeats adjacent to CTCF sites facing the regulated promoters. Distal silencers directly modulating transcription at spatially proximal promoters thus appears to be a previously overlooked and prevalent regulatory element. We propose that ancient transposable elements, which are often systematically repressed as a genomic defense mechanism, may be co-opted for endogenous gene regulation.

Results

An extensive promoter interactome comprises constitutive and cell type-specific contacts in mouse thymocyte development

To explore to what extent promoter-mediated interactions are remodeled in response to developmental cues, we performed Promoter Capture coupled to *in situ* Hi-C (PCHi-C) (Schoenfelder et al., 2015a) in mouse $CD4^-$ $CD8^-$ $CD44^-$ $CD25^+$ (double negative; DN3) and $CD4^+$ $CD8^+$ (double positive; DP) thymocytes. These represent cell populations just before and after the checkpoint for productive rearrangement of the T cell receptor-β gene, which is essential for generating productive T cells and is accompanied by well-characterized transcriptional and epigenomic changes at hundreds of gene loci (Egawa and Littman, 2011; Koch et al., 2011; Pekowska et al., 2011; Zhang et al., 2012). For comparison with a completely unrelated cell type, we also re-analyzed PCHi-C data generated in mouse embryonic stem (ES) cells (Schoenfelder et al., 2015a). Using a custom method which identifies peaks of locally increased PCHi-C signal, and is more robust to individual restriction fragment variations across biological replicates (see Methods), we identified thousands (38,399 DN3; 34,603 DP; 24,487 ES; **Fig 1a** and **Table S1**) of promoter-centered interactions. As previously reported (Schoenfelder et al., 2015a; Mifsud et al., 2015), these interacting regions are predominantly contained within topologically associated domains (TADs) (e.g. 80% for ES), consistent with their proposed role in delimiting the functional

range of gene regulatory elements (Symmons et al., 2014; Lupianez et al., 2015). As may be expected of more decondensed chromatin loci, active genes participate in an overall greater number of interactions, which span a larger genomic distance ($p < 2x10^{-16}$; Wilcoxon rank sum test, in each case; **Fig 1b,c** and **Fig S1**). Globally, the interacting regions are highly enriched in binding of CTCF and cohesin (**Fig 1d,e** and **Fig S1**), consistent with these factors' well established role in chromatin looping (Splinter et al., 2006; Hadjur et al., 2009; Kagey et al., 2010). Recently, it has been shown that CTCF-mediated loops predominantly form between binding sites with convergent motifs (Rao et al., 2014; de Wit et al., 2015; Guo et al., 2015). Although facing CTCF-bound sites are only found at both the promoter and interacting region in a minority of cases (3.8% in DN3; 11.4% in DP), the CTCF motifs at non-promoter regions exhibit a strong bias towards the orientation facing towards the interacting gene (65.6% in DN3, $p = 9x10^{-249}$; 58.5% in DP, $p = 3x10^{-148}$; binomial distribution). CTCF orientation thus appears to influence gene interactions, without necessarily participating in CTCF-CTCF interactions directly at the promoter. The interacting loci are also enriched in histone modifications associated with transcriptional regulation, such as H3K4me1, H3K27ac and H3K27me3, as well as binding of RNA polymerase II (PolII) and hematopoietic/thymocyte transcription factors, suggesting that many are putative distal regulatory elements, particularly enhancers (**Fig 1d,e**). In line with this, PCHi-C detected thymocyte-specific interactions between characterized distal thymocyte enhancers and their target genes, such as *Ikzf1*, *Bcl11b* (Isoda et al., 2017; L. Li et al., 2013) and *Satb1* (**Fig 1f,g** and **Fig S2**). When comparing the repertoires of called promoter-centered contacts, a high degree of cell type specificity is apparent, even for the closely-related thymocyte populations (**Fig 1a**; Jaccard index 0.32). Pluripotent ES cells share nearly two-fold fewer interactions with either thymocyte type (Jaccard indices 0.20 and 0.18, with DN3 and DP respectively), consistent with reports of related cell types having more similar chromosome topologies

(Javierre et al., 2016). However, a core (10.7%) of stable interactions is maintained in all three cell types studied. As may be expected, the regions participating in stable interactions were even more highly enriched in CTCF binding sites that are conserved across these cell types (**Fig S1**). We validated several maintained and cell type-specific interactions by 4C-seq (**Fig 1h** and **Figs S2,S3**), confirming that these differences cannot be attributed to any technical issues with bait capture or PCHi-C analysis. Overall, the conclusions from a general analysis of mouse thymocyte PCHi-C data are consistent with those obtained from other studied differentiation systems: a complex network of constitutive and cell type-specific promoter-centered interactions, many with CTCF sites and/or putative enhancers.

Dynamic promoter interactions correlate with gene repression as well as activation

To explore the developmental dynamics of promoter-centered topologies in a more quantitative manner, we compared quantile-normalized PCHi-C scores between DN3 and DP datasets for called interactions with one or the other cell type. We reasoned that plotting such interaction differences against transcriptional output of the corresponding gene (from RNA-seq data) would resolve "instructive" promoter-enhancer loops (interaction increases concomitantly with transcriptional increase) from "permissive" or "poised" ones (interaction is present in both cells, and does not change, but gene expression increases). When analyzing DN3 or DP promoter interactions, we indeed observed hundreds of cases of both instructive and permissive promoter-enhancer loops (**Fig 2a-d**; **Table S1**).

However, perhaps surprisingly, an equivalent number of interactions that are increased in one cell type are associated with a transcriptional *decrease* of the contacted gene (**Fig 2a,e**). These could represent promoter interactions with distal silencer elements, and/or setting up of new poised enhancers which will activate gene expression at later developmental stages. A large number of cell type-specific interactions also appear unrelated to any transcriptional change of the associated genes, implying that dynamic chromatin architectures can also be uncoupled from underlying gene activity (**Fig 2a,h**). We operationally classed promoter interactions based on these behaviors (**Fig 2a**): A (putative *active enhancers*; interaction increases, transcription increases; **Fig 2c**); B (putative *poised enhancers*; interaction unchanged, transcription increases; **Fig 2d**); C (putative *cell type-specific silencers*; interaction increases, transcription decreases; **Fig 2e**); D (interaction unchanged, transcription decreases; **Fig 2f**); E (interaction unchanged, transcription unchanged; **Fig 2g**); F (interaction increases, transcription unchanged; **Fig 2h**). The last two classes can be subdivided into those where the target genes are silent (E_s, F_s) or active (E_a, F_a) in both thymocyte populations. We observed similar behavior when comparing DN3 or DP interactions with ES cells, and *vice versa*, suggesting that such putative cell type-specific silencer interactions are not limited to thymocyte lineages (**Fig S4**), and several of these interactions were validated by 4C-seq (**Figs S2** and **S3**). As expected of putative enhancers, class A and B interacting regions were enriched in H3K27ac, H3K4me1 and PolII, and depleted in repressive histone marks like H3K27me3 and H3K9me3 (**Fig 2b**). The same profile was observed for class D interactions, implying that these represent developmentally stable enhancer interactions with genes that are actually upregulated in the other thymocyte population. Conversely, E_s and F_s class interactions with stable silent genes were enriched in repressive marks and depleted with active marks. Overall, a more quantitative comparative analysis of PCHi-C datasets revealed dynamic and stable promoter interactions that are not just linked to transcriptional activation,

but also to more complex situations that may be linked to transcriptional repression, poising, or otherwise unrelated to transcription.

Both active and poised enhancers participate in interactions with target genes

To validate the putative enhancers identified from our promoter CHi-C results in DN3 and DP cells, we first compared them with STARR-seq (self-transcribing active regulatory region sequencing) data obtained from the immature thymocyte cell line, P5424, which has a transcriptome profile in between that of DN3 and DP (closer to DN3) (Vanhille et al., 2015). STARR-seq comprises the high-throughput assessment of functional enhancer activity of DNA elements transfected within libraries of episomal constructs (Arnold et al., 2013), and was applied to 7152 DNase hypersensitive sites in P5424 cells (Vanhille et al., 2015). When comparing with the most similar DN3 cells, around half of the sites classed as having “strong” (218/433) or “weak” (1115/2279) enhancer activity were found to interact with promoters, although these formed a small proportion of the entire DN3 promoter interactome. Notably, the class A interactions had proportionally nearly two-fold more “strong” enhancers than other interaction classes, whereas “weak” enhancers were more prevalent in class B interactions (Fig. 3a), implying that the strongest enhancers may be the most cell-type specific. We further validated some thymocyte-conserved and DN3-specific putative enhancers (based on their interaction dynamics) in luciferase reporter assays in P5424 and ES cells (Fig. 3b). As expected, reporter expression was highly upregulated in the thymocyte lineage but minimally affected in ES cells. A DN3-specific, strong enhancer of particular interest was located ~1.3 Mb downstream of the proto-oncogene Myc; the interaction with this gene was specific to DN3 cells in both CHi-C and 4C-seq, correlating with gain of H3K27ac at the enhancer (Fig. 3b-e).

This enhancer is conserved in humans and was previously found to be duplicated in many T acute lymphoblastic leukemias (T-ALL) via overexpression of *MYC* in T cell precursors (Herranz et al., 2014). Deletion of the enhancer in mice also caused defects in mature T cell production via a block at the DN3-to-DP transition (Herranz et al., 2014). To obtain proof of principle for this enhancer as a potential T-ALL drug target, we used the nuclease-dead Cas9 system to recruit the transcriptional repressor KRAB to the distal *Myc* enhancer, resulting in a ~5-fold reduction of *Myc* expression (**Fig. 3f**).

Beyond the clues given by comparison with the STARR-seq data, we were unable to obtain much information on what could distinguish instructive from permissive enhancer looping models. Extensive clustering analysis (data not shown) did not provide any new insight that was not already obtained from the more global analyses already presented (**Fig 2a,b**): there appear to be an equivalent number of "poised" and active enhancer-promoter looping interactions, and there are no apparent differences in epigenetic marks between enhancers participating in class A or B interactions. Similarly, no thymocyte transcription factor combinations come out of clustering analyses as driving any particular class of enhancer interaction.

Distal promoter-interacting silencers are prevalent in the mouse genome

We next looked closer at the interacting elements which correlated with target gene repression. Whereas some classes of these interactions (E_S and F_S) were clearly depleted for active histone marks and weakly enriched in repressive histone marks, the "dynamic" list of putative silencers from class C demonstrated some enrichment for both active and repressed marks (**Fig 2b**).

When looking in more detail at the class C interactions by hierarchical clustering (data not shown), a subset of around a quarter of interactions do indeed resemble enhancers, with clear enrichment for H3K27ac and depletion for repressive marks. We posit that these represent poised enhancers for genes which are upregulated in other thymocyte lineages earlier and/or later than the DN3-to-DP transition which we interrogated. Nevertheless, a much greater proportion of interacting regions than for class A and B interactions were devoid of any known histone modifications, suggesting that they may indeed be enriched in a different class of regulatory element. To functionally assess whether putative silencer elements from the C and E_S class do indeed confer intrinsic transcriptional repression, we performed luciferase reporter assays in P5424 and ES cells with candidate regions inserted upstream of the reporter under control of a strong SV40 enhancer (**Fig 4a**). To avoid confounding technical problems from using large (2 kb) inserts, these results were normalized to those from equal-sized "neutral" inserts, which were selected and tested in each cell type to not induce significant up- or downregulation of a reporter under the control of either a minimal promoter or SV40 enhancer. Most tested silencers caused significant reporter repression in both P5424 and ES cells, implying that these elements may be *bona fide* distal silencers, and that their intrinsic effects on transcription are less cell type-specific than enhancers, perhaps recruiting ubiquitous factors However, a DN3-specific region interacting with the proto-oncogene *Dek* only conferred efficient reporter silencing in P5424 cells, suggesting that some tissue specificity can be present. Although these regions were not particularly enriched in known repressive chromatin modifications, such as K3K9me3 or H3K27me3, these may be underestimated by reliance of the CHi-C analysis on sequences mapping to unique, non-repetitive regions, which are enriched in those marks. Similarly, we found no enrichments for motifs of known transcription repressors, such as REST, but such analyses are technically hampered by limited resolution of the CHi-C interacting regions to single restriction

fragments at best, and the role of such factors cannot be discounted. As for all promoter-interacting regions, we frequently found candidate distal silencers to contain a CTCF motif facing the targeted promoter, and noted that many of these were juxtaposed to repetitive DNA elements derived from ancient transposable elements (TEs). In general, TEs and their diverged, non-transposing variants, are transcriptionally shut down by a variety of mechanisms, especially in the germline (Friedli & Trono, 2015). However, particular long terminal repeats (LTRs) and short interspersed nuclear elements (SINEs), which initially activated transcription of ancient viral or transposable elements, have been described to be evolutionarily adopted ("exapted") as enhancers of cellular genes (Bejerano et al., 2006; Lowe, Bejerano, & Haussler, 2007; Sasaki et al., 2008). Exploring the link between ancient TEs and putative silencers further, we found that CTCF-linked interactions with silent genes were significantly depleted in SINEs and enriched in juxtaposed long interspersed nuclear elements (LINEs); LTRs were neither enriched or depleted (**Fig 4b**). We therefore reasoned that the transcriptional repressive mechanisms intrinsically brought to TEs (particularly LINEs) by the cellular host defenses could also affect transcription of endogenous genes, if brought to them via CTCF-mediated chromatin looping. In support of this hypothesis, luciferase reporter assays with just the TE component of the previously validated silencers gave equal or better transcriptional repression in most cases (**Fig 4c**).

To provide further support for the model that TEs can generally repress distal promoters when brought into their spatial proximity, we are currently using CRISPR/Cas9 technology to specifically delete the TEs of these luciferase assay-validated silencers in P5424 and ES cells, in parallel with deletion of their juxtaposing CTCF sites. We predict that deletion of the CTCF site may perturb looping between the TE and the gene, which will be tested by 4C-seq, and cause derepression of the target gene, which will be tested by qRT-PCR. Specific deletion of the TE may be expected to cause the same derepression without

affecting chromatin topology (**Fig S5a**). The large functional redundancy among enhancers, whereby deletion of one enhancer has only minor phenotypic effects to confer evolutionary robustness of developmental gene regulation, is becoming recently appreciated (Osterwalder et al., 2018). Such derepression effects, if any, of these putative silencer deletions could potentially be small. Finally, we have preliminary evidence by 4C-seq that artificial induction of the *Bcl6* gene in ES cells (see TAD capture results chapter) may perturb chromatin looping to luciferase assay-validated silencers, which form interactions with the promoter in DN3 and ES cells, but not DP cells where *Bcl6* is highly expressed (**Fig S5b**). Thus although the mechanism remains unclear, escape from such inhibitory chromatin loops may be an aspect of developmental gene activation.

Discussion

Recent promoter CHi-C studies have assigned putative enhancers to target genes (Hughes et al.. 2014; Schoenfelder et al., 2015a; Sahlen et al., 2015), and assessed the dynamics of promoter-enhancer loops during different differentiation or developmental models, coming up with varying conclusions of highly dynamic (Siersbaek et al., 2017) or a complex mixture of dynamic and more stable chromatin interactions (Freire-Pritchett et al., 2017; Rubin et al., 2017). In our study we come to the similar conclusion that thymocyte differentiation involves the interplay of both permissive and instructive promoter-enhancer contact networks. Whereas one study claimed to identify the specific transcription factors distinguishing permissive from instructive loops during epidermal differentiation (Rubin et al., 2017), such a simple model was not apparent in thymopoesis. It is known that the cocktail of transcription factors driving thymic differentiation, such as Ikzf1, Runx1, GATA3, Bcl11b and Tcf12, do not have expression patterns limited to specific stages, but instead regulate different networks of genes with much temporal complexity (Thompson & Zúñiga-Pflücker, 2011). Therefore the

regulation of each gene is likely to take place within a context of chromatin accessibility, prior chromosome topology and dynamic histone modifications, confounding a one-size-fits-all model. With the exception of identifying one isolated promoter interaction linked to repression (Mifsud et al., 2015) or a complementary study of Polycomb-mediated promoter-promoter interactions (Schoenfelder et al., 2015b), the previous CHi-C studies made no assessment of chromatin topologies correlated with gene downregulation. However, we observe not just in thymocytes, but when comparing thymocytes with ES cells, a large population of promoter-centered interactions linked to target gene repression (**Fig 2a, Fig S4**). The major reason these are likely to have been overlooked is that the relatively limited resolution of CHi-C (single restriction fragments) compared to ChIP-seq means that potentially functional chromatin interactions are easiest homed in with a characteristic histone modification peak (e.g. H3K27ac) or transcription factor binding site. Thus whereas putative enhancers are relatively easy to characterize, insufficient information on the epigenetic signature (if any exists) for silencers means that they are much more difficult to define precisely. Within these limitations, we have uncovered a possible role for TEs, particularly LINEs, to indirectly repress gene transcription when brought by a loop to a target promoter. As the majority of intrinsic silencer sequences we have uncovered appear to be non-cell-type-specific (or at least conserved between ES and thymocyte lineages), we speculate that if these are indeed exapted for developmental gene regulation, then it will likely do so at the level of chromatin topology. In support, preliminary data suggest that such loops are lost or reduced on transcriptional induction, although it remains to be seen if transcription is a cause or consequence of such loop remodeling at endogenous developmental gene loci, nor how the remodeling can be brought about. One study implicated the histone variant H2A.Z in repressive loop formation at a specific gene locus (Dalvai et al., 2013), but the generality of this has yet to be assessed.

Overall, chromatin looping appears highly developmentally dynamic. Although most efforts to date have concentrated on interactions with enhancers linked to transcriptional activation, these appear to be the tip of the iceberg of the full promoter interactome: just as many interactions correlate with gene downregulation, and even more are not linked to expression changes of the targeted gene at all. It remains a daunting challenge to tease apart just how many of these promoter interactions, while robust and reproducible, are actually frequent in a cell population *and* functionally important. There is unlikely to be a simple mechanism that explains all topology phenomena. For example, some category F interactions appear to involve the promoter swapping contacts with one flanking TAD for those with the other TAD, reminiscent of the shift in regulatory "archipelagoes" described at the Hox loci (Andrey et al., 2013), but this is just a small minority (**Fig 2h**). Only once the functionally relevant non-enhancer-linked chromatin topologies start to be mechanistically teased apart, will we gain a full understanding of if and how chromosome folding controls the genome.

Materials and Methods

Isolation of mouse DN3 and DP thymocytes

Thymuses were dissected from 6-8 week old c57/Bl6 mice, and DN3 and DP cell populations were purified by fluorescent assisted cell sorting (FACS), following the protocol of (Oravecz et al., 2015).

Hi-C and promoter CHi-C

In situ Hi-C was performed with HindIII, essentially as in (Vietri Rudan, Hadjur, & Sexton, 2017). Promoter capture was performed with the same oligonucleotide design and methodology as in (Schoenfelder et al., 2015a).

4C-seq

4C-seq was performed as in (Noordermeer et al., 2014) and analysed as in de (Wit et al., 2015).

P5424 and ES cell culture and transfection

P5424 cells were maintained as in (Vanhille et al., 2015); ES cells as in (Bibel et al., 2007).

Luciferase reporter assays

Luciferase assay constructs were cloned into pGL3 variants and subjected to double luciferase assays (with a Renilla construct as a transfection control), essentially as in (Mifsud et al., 2015).

dCas9-KRAB experiments

P5424 cells were co-transfected with a dCas9-VP64 (Addgene) and custom-made four-guide RNA vectors (generated by the IGBMC platform), sorted by FACS for GFP expression to obtain the most highly transfected cells, and then cDNA was harvested after 48 hr to test for target gene expression by qRT-PCR.

Hi-C read processing and filtering

Hi-C reads were pre-processed and valid reads filtered following a pipeline very similar to that of Sexton et al. (2012). See also the previous results chapter for Hi-C quality controls.

CHi-C interaction calling and quantile normalization

CHi-C interactions were called by PromoMaxima (see accompanying manuscript), using the following parameters: (-w =50; -s=0.05, -d=30000). For comparison of different datasets, the loess smoothed profiles were quantile normalized by the limma package in R.

Epigenomic profile sources and pre-processing

ChIP-seq data were downloaded from GEO database for these histone marks: H3K4me1, H3K27ac, H3K4me3, H3K27me3 and for these transcription factors: Ikaros, Runx1, Ets1, GATA3 and for PolII, CTCF and cohesin (**Table S2**). I aligned ChIP-seq on mm9 genome using bowtie2. Then, I called peaks using Erange V4.0. In Erange, mapped reads in the SAM file are first transformed into native Erange reads store .rds file. Then, peaks are identified with the peaks finder tools in Erange with –nodirectionality and –notrim parameters. Erange returns a per-peaks p-value. Wig files of Histone marks and transcription factors were made using the makewiggle.py script from rds files with 20 bp coverage. Then, Wig files are quantile normalized between different cell types for each histone mark or transcription factor supposing that the antibody efficiency is the same for different cell types.

All GEO datasets except some data with SoliD reads were similarly processed. For SoliD data (Ikaros, Runx1 and PolII), peak files and wig files were downloaded then, binned into 20 bp and quantile normalized.

RNA-seq data for DN3, DP and ES cells from GEO database (**Table S2**) were obtained as fastq files. Reads were mapped to mm9 genome using bowtie2. Mapped reads in SAM file are then transformed into rds file by using Erange V4.0 tools (makerdsfrombowtie.py). For each gene, Erange counts unique reads falling on the gene models using rpkm normalization. The

output is a text file with each line corresponding to a specific gene with its corresponding rpkm value.

Computation of enrichment for epigenetic marks

The enrichment for chromatin marks and transcription factor in interacting fragments was calculated using the proportion of fragments that overlap with a peaks for the mark state or transcription factor, divided by the proportion of all non-bait fragments that overlap with such a peak. Then, resulting values were converted to its log2 value, so that positive values represent an enrichment compared with all non-bait fragments and a negative value represents depletion.

To assign interacting fragments to an expression class, the interacting fragment must interacts with baits from the same expression class otherwise it is excluded from the enrichment analyses.

Clustering analysis

For each interacting fragment, the enrichment of histone mark or transcription factor is calculated as follow:

- Each interacting fragments score is calculated as the mean of overlapping Histone marks or transcription factors
- The fold change enrichment corresponds to the ratio of the score of interacting fragment to the score of all restriction fragments

All interacting fragments are then clustered using Euclidian distance and Ward.D method based on their corresponding fold change enrichment value.

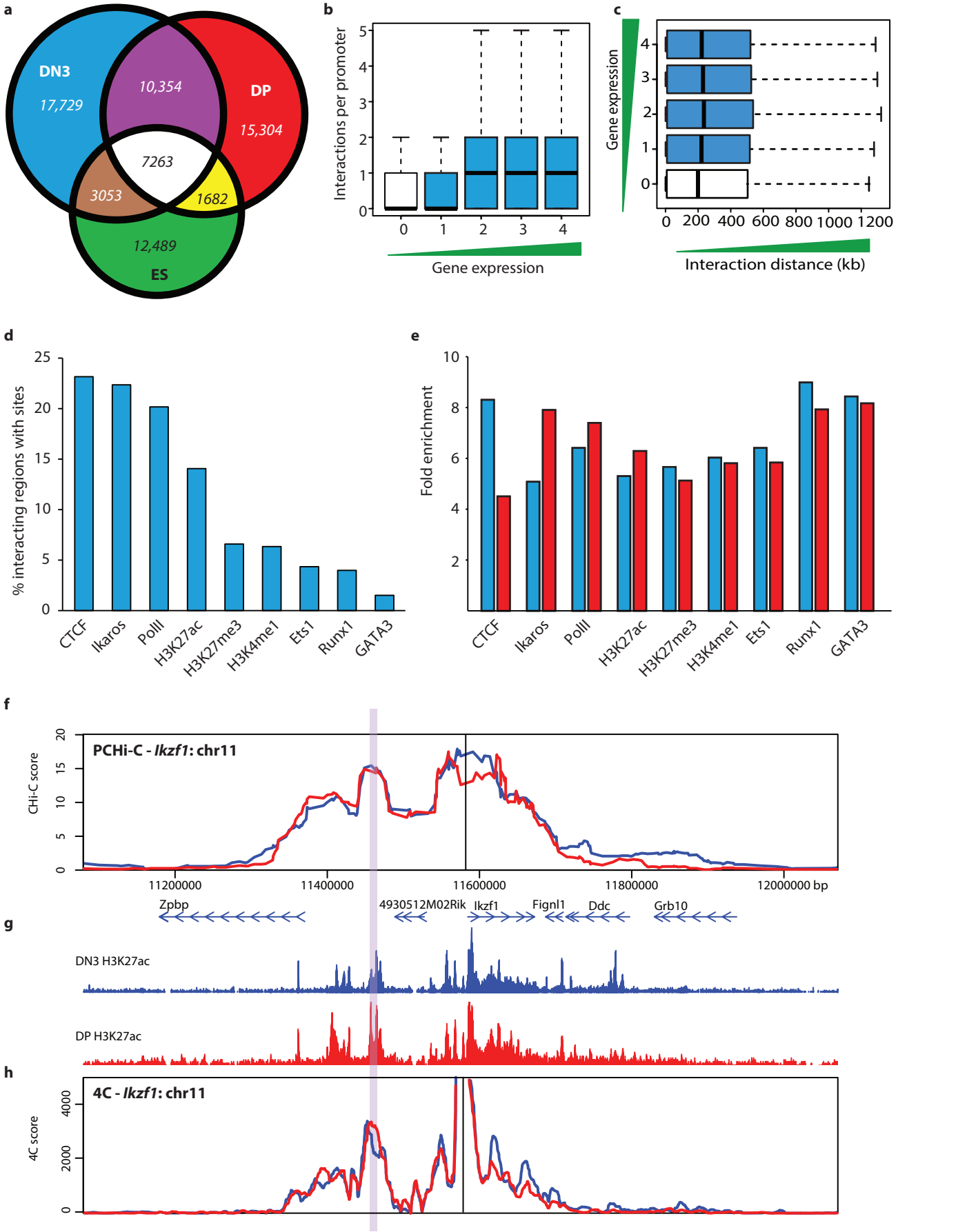


Fig 1. The mouse thymocyte promoter interactome.
a) Venn diagram of called promoter interactions in DN3 (blue), DP (red) and ES (green) cells.
b) Distribution of numbers of DN3 promoter-centered interactions, classed by gene expression of target gene.
c) Distribution of DN3 promoter-centered interaction distances, classed by gene expression of target gene.
d) Proportions of DN3 promoter-centered interactions containing peaks for different histone modifications or bound factors.
e) Relative enrichment of DN3 (blue) and DP (red) promoter-interacting regions for various histone marks or bound factors.
f) CHi-C profile (DN3 blue, DP red) for local interactions with Ikzf1 gene. g) DN3 (blue) and DP (red) H3K27ac ChIP-seq profile for the same genomic region.
h) 4C profile for interactions with the Ikzf1 promoter in DN3 (blue) and DP (red) cells. A called interaction between Ikzf1 and a putative enhancer in both DN3 and DP cells is denoted by a purple stripe.

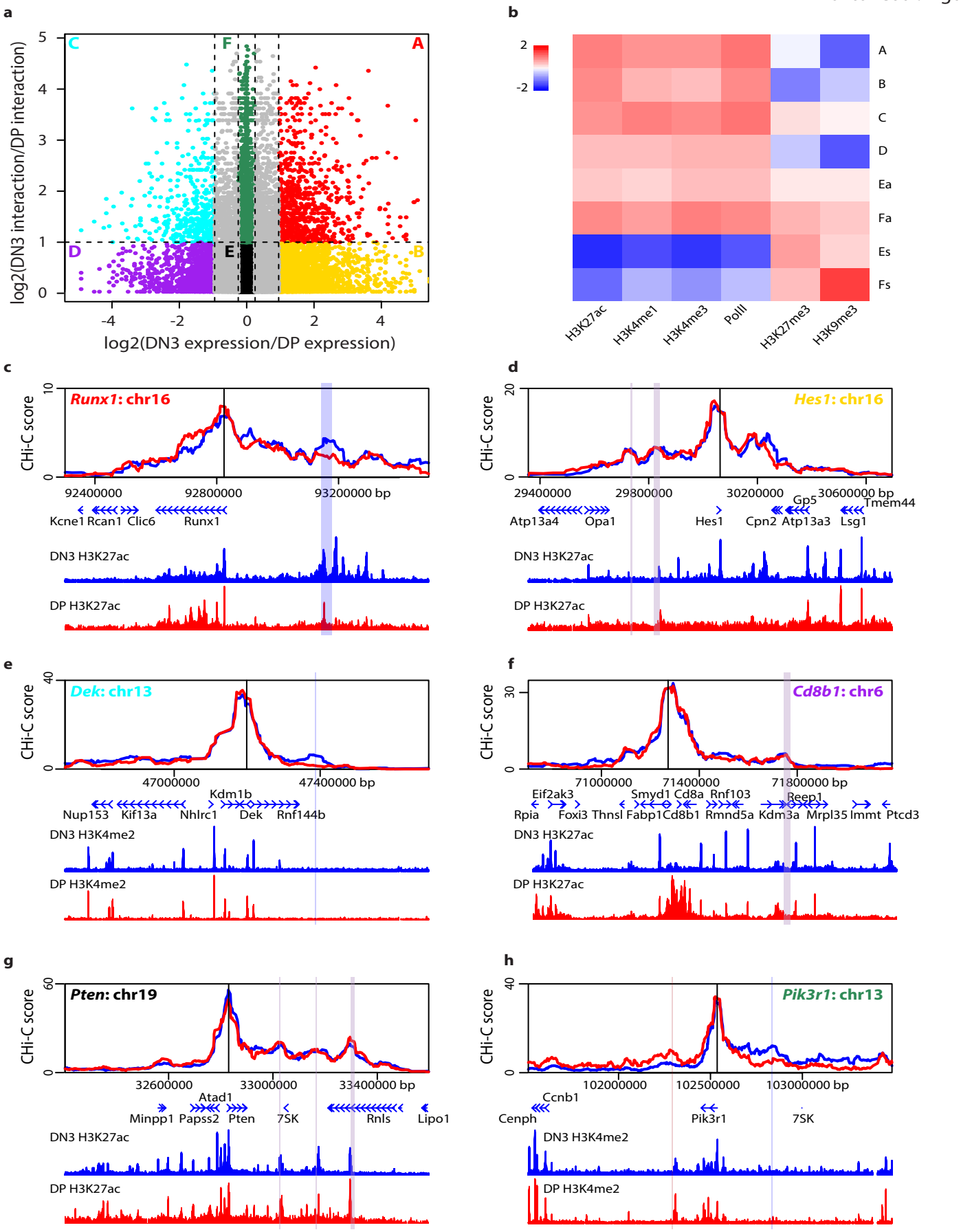


Fig 2. Stable and dynamic promoter interactions linked to transcriptional activation and repression.
a) Scatter plot for all interactions called with DN3 promoters, plotting difference in CHi-C interaction score between DN3 and DP against difference in gene expression between DN3 and DP, as computed from RNA-seq results.
Different classes of interactions are labeled in different colors: A (red), B (gold), C (cyan), D (purple), E (black), F (dark green).
b) Heat map showing relative enrichment or depletion (on log2 scale) of different histone marks and bound factors in regions corresponding to DP interactions of different classes, called as in a, except that classes E and F are further categorized into those with active (Ea, Fa) or silent (Es, Fs) genes. c-h) CHi-C screenshots (DN3 in blue; DP in red) for DN3 interactions of different classes: c) A with Runx1; d) B with Hes1; e) C with Dek; f) D with Cd8b1; g) E with Pten; h) F with Pik3r1. ChIP-seq profiles for H3K27ac or H3K4me2 are shown alongside (DN3 in blue; DP in red). Different colored stripes indicate different interactions (blue DN3-specific; red DP-specific; purple conserved in both thymocytes).

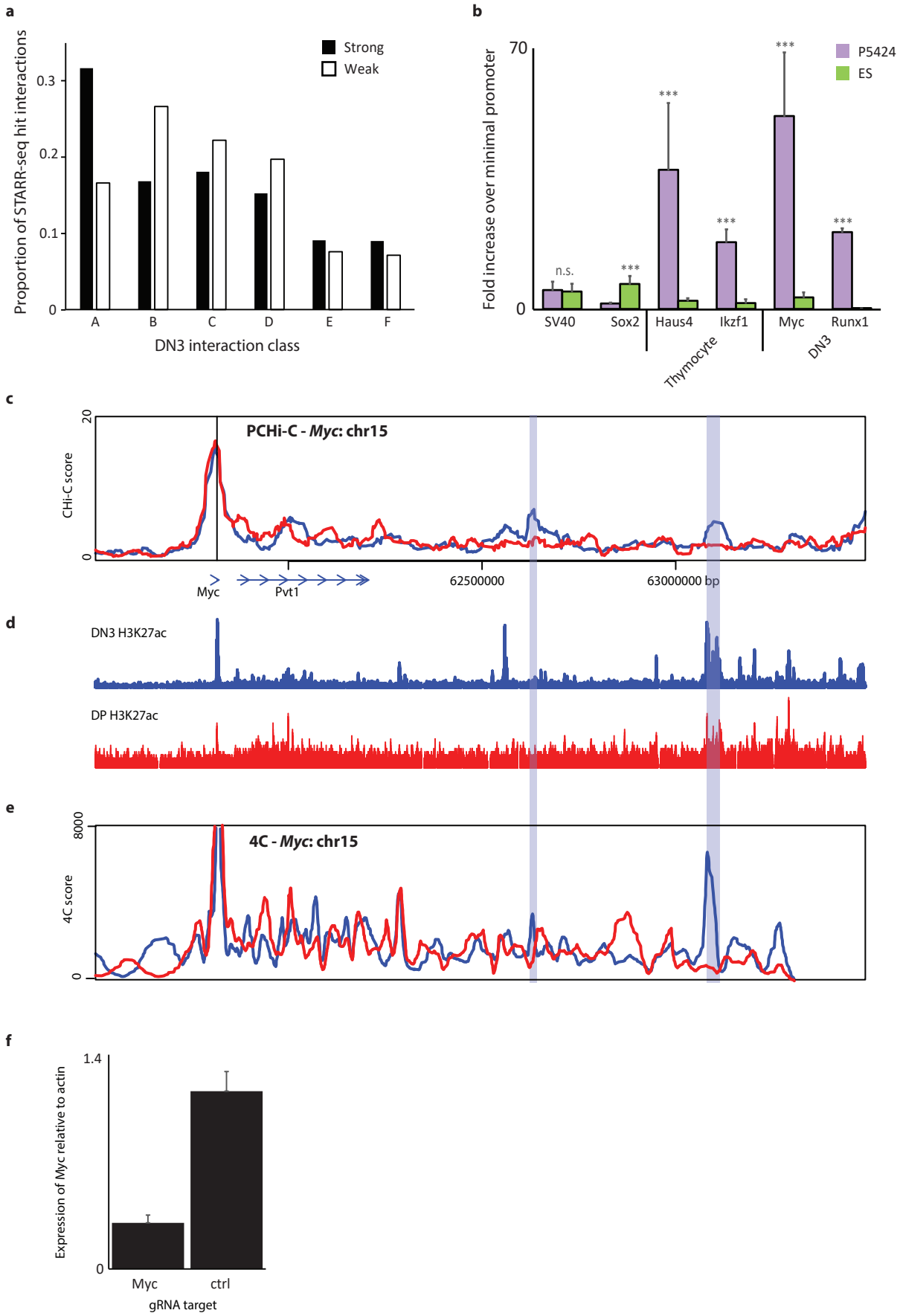


Fig 3. Thymocyte-specific and dynamic enhancers.
a) Proportion of total strong (black) and weak (white) P5424 STARR-seq hits (from Vanhille et al., 2015) present in DN3 promoter interactions, classed according to interaction type.
b) Luciferase reporter assay results in P5424 (purple) and ES cells (green), expressed as fold increase in reporter expression over minimal reporter constructs. Results are shown for a constitutive (SV40), and ES-specific (Sox2) controls, as well as CHi-C-called thymocyte- specific and DN3-specific interacting enhancers. *** $P < 0.001$; two-tailed t-test comparing the two cell types, with Benjamini-Hochberg multiple testing correction. c) CHi-C profile (blue DN3, red DP) around the Myc gene. d) H3K27ac ChIP-seq profiles (blue DN3, red DP) around the same region. e) 4C profile for interactions with the Myc promoter (blue DN3, red DP).
f) qRT-PCR results for Myc expression, expressed relative to actin, in P5424 cells after treatment with dCas9-KRAB and guide RNAs directed to either the Myc enhancer or an unrelated genomic region as control.

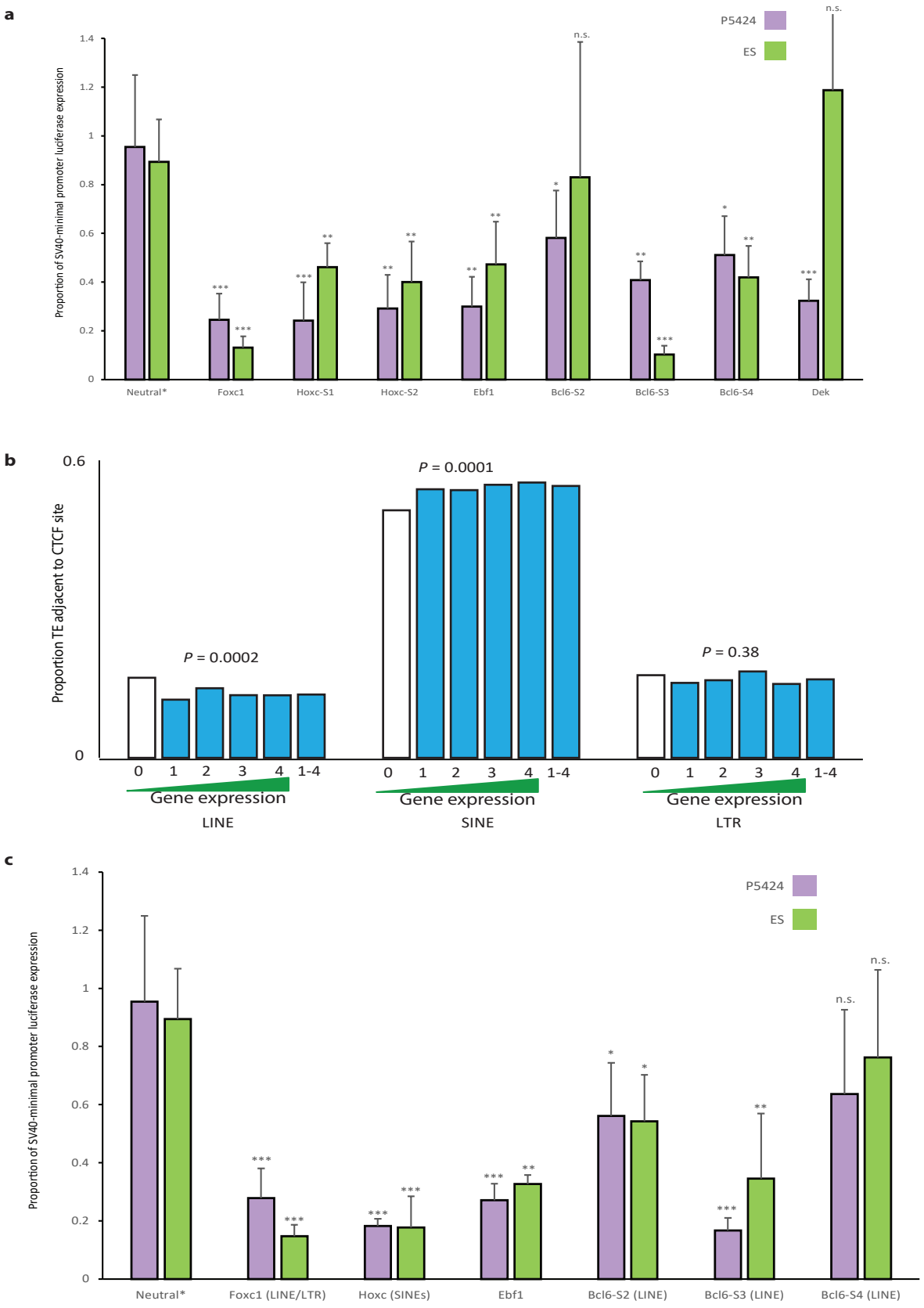


Fig 4. Distal silencers may regulate contacted genes.
a) Luciferase reporter assays for 2 kb test inserts upstream of the SV40 enhancer/promoter in plasmids transfected in P5424 (purple) or ES (green) cells. Reporter expression is expressed as proportion of the SV40 enhancer/promoter construct without other insert. P-values are calculated by two-tailed t-tests comparing the test insert with its corres--ponding cell type- matched neutral sequence, with Benjamini Hochberg multiple testing correction. *** P < 0.001; ** P < 0.005; * P < 0.05 b) Proportion of interacting regions containing a particular class of TE adjacent to a CTCF mo--tif, classed according to expression of the interacting gene. P-values are given from the Fisher's exact test, comparing the silent gene interactions with all active gene interactions.
c) Luciferase reporter assays for ~500 bp test inserts upstream of the SV40 enhancer/promoter in plasmids transfected in P5424 (purple) or ES (green) cells, as in

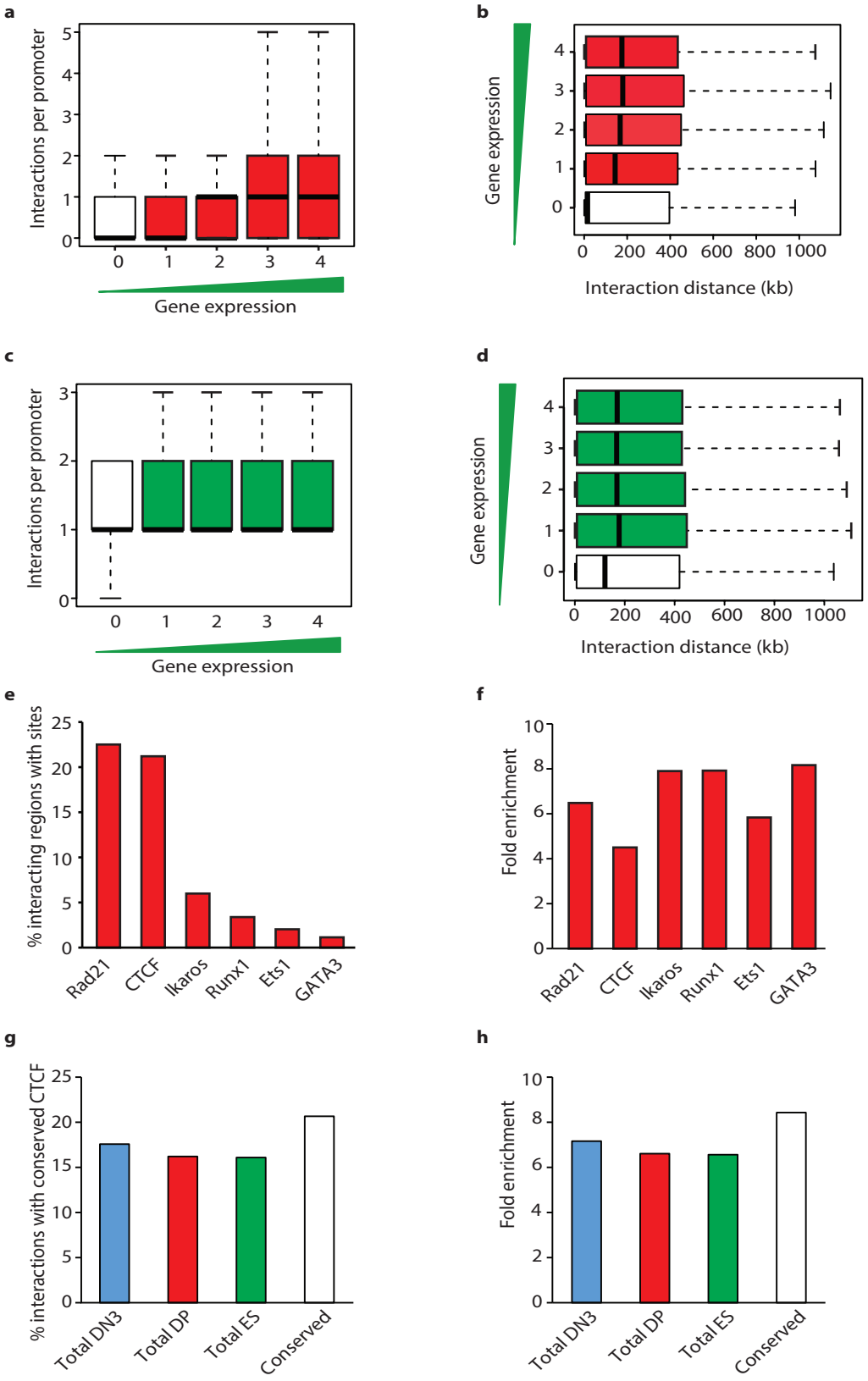


Fig S1. The mouse thymocyte and ES promoter interactome features.
a) Distribution of numbers of DP promoter-centered interactions, classed by gene expression of target gene.
b) Distribution of DP promoter-centered interaction distances, classed by gene expression of target gene.
c) Distribution of numbers of ES promoter-centered interactions, classed by gene expression of target gene.
d) Distribution of ES promoter-centered interaction distances, classed by gene expression of target gene.
e) Proportions of DP promoter-centered interactions containing peaks for different histone modifications or bound factors.
f) Relative enrichment of DP promoter-interacting regions for selected histone marks or bound factors. g) Total percentage and
h) relative enrichment of different cell type promoter interactions, and interactions conserved in all cell types, for conserved CTCF sites.

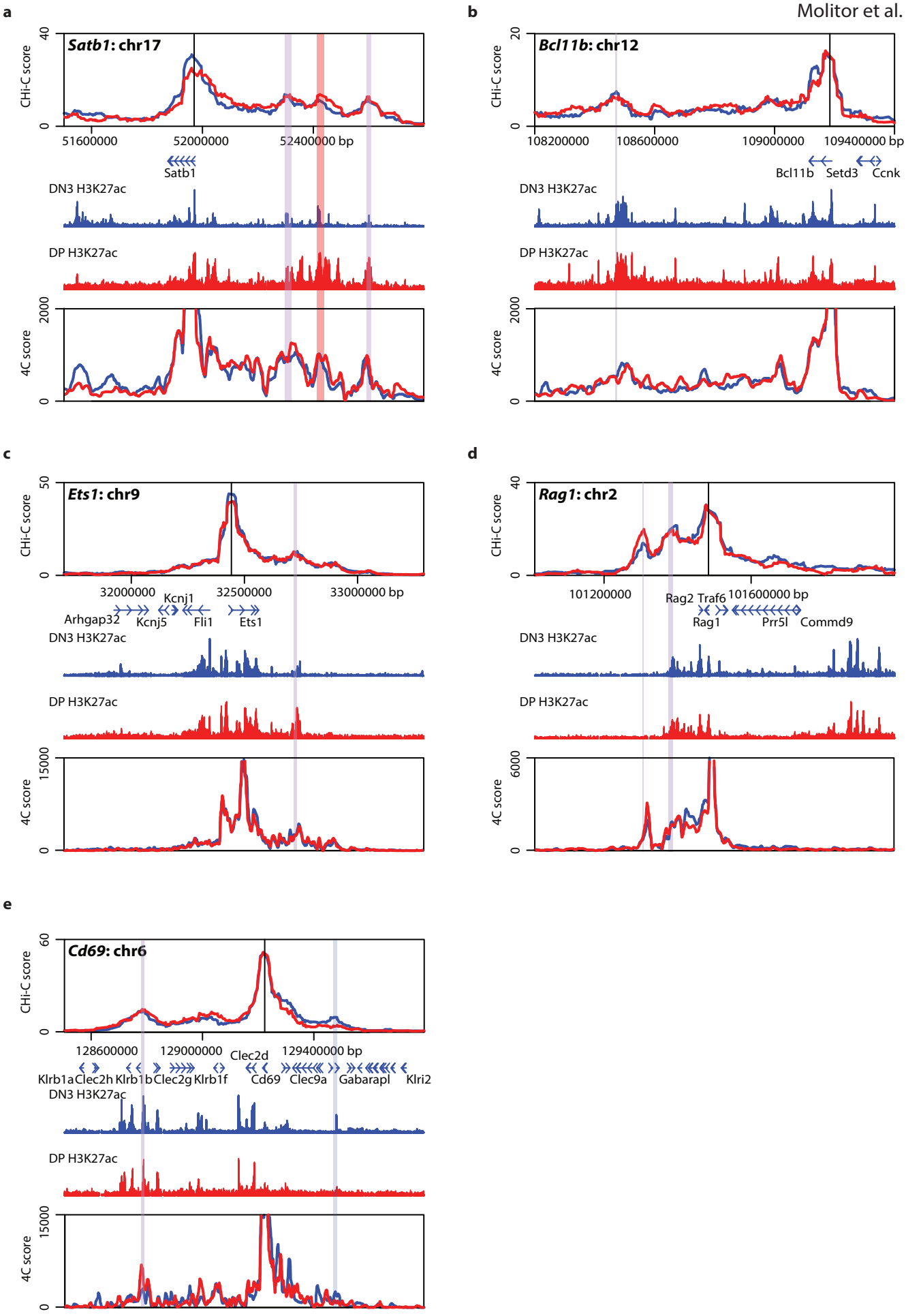


Fig S2. Stable thymocyte promoter interactions.

a-e) Selected CHi-C and corresponding 4C profiles for interactions which are predominantly conserved in DN3 and DP cells. Called interactions are given by purple (conserved), red (DP) or blue (DN3) stripes.

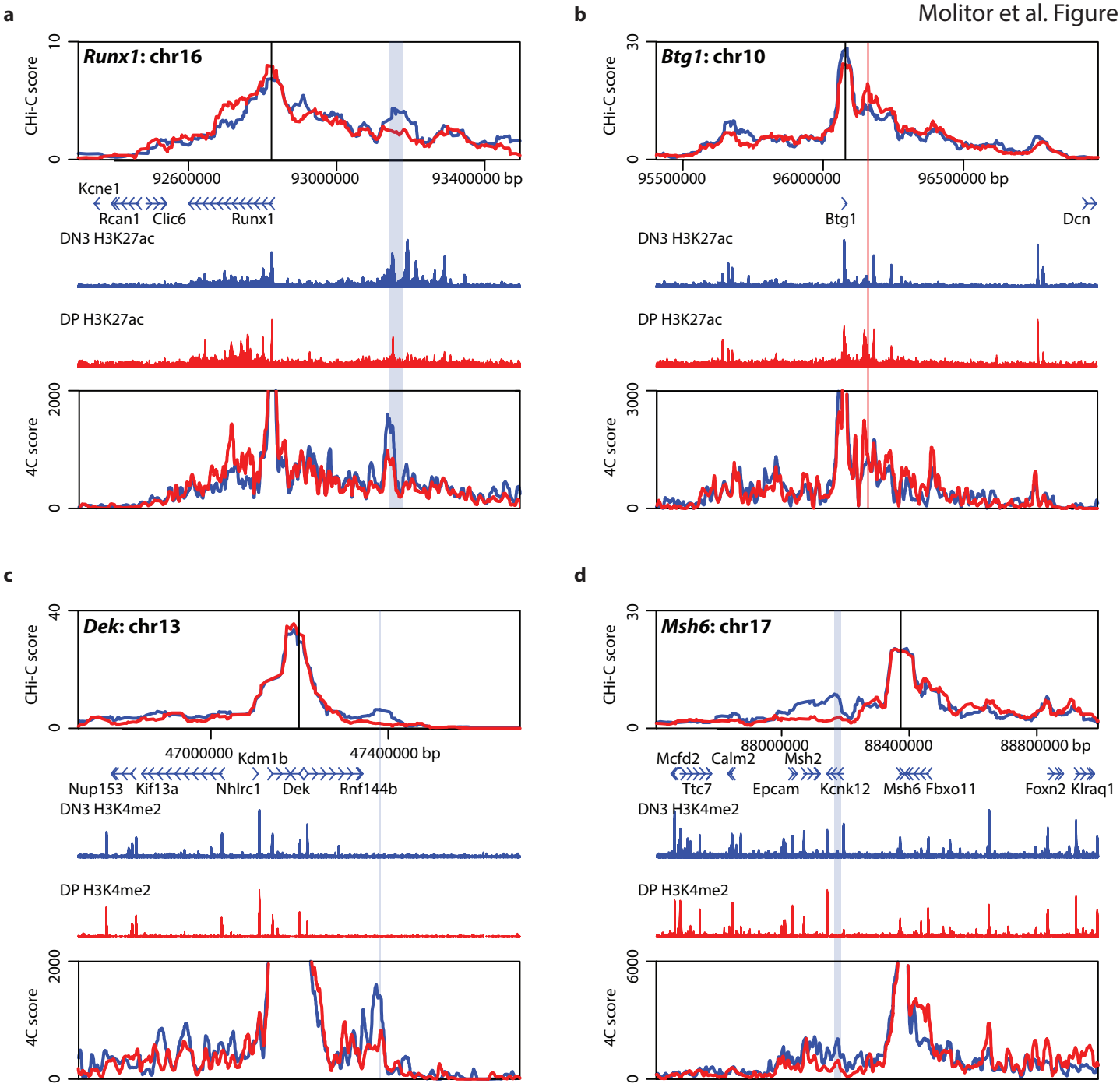


Fig S3. Dynamic thymocyte promoter interactions.

a-d) Selected CHi-C and corresponding 4C profiles for interactions which are cell type- specific. Called interactions are given by red (DP) or blue (DN3) stripes.

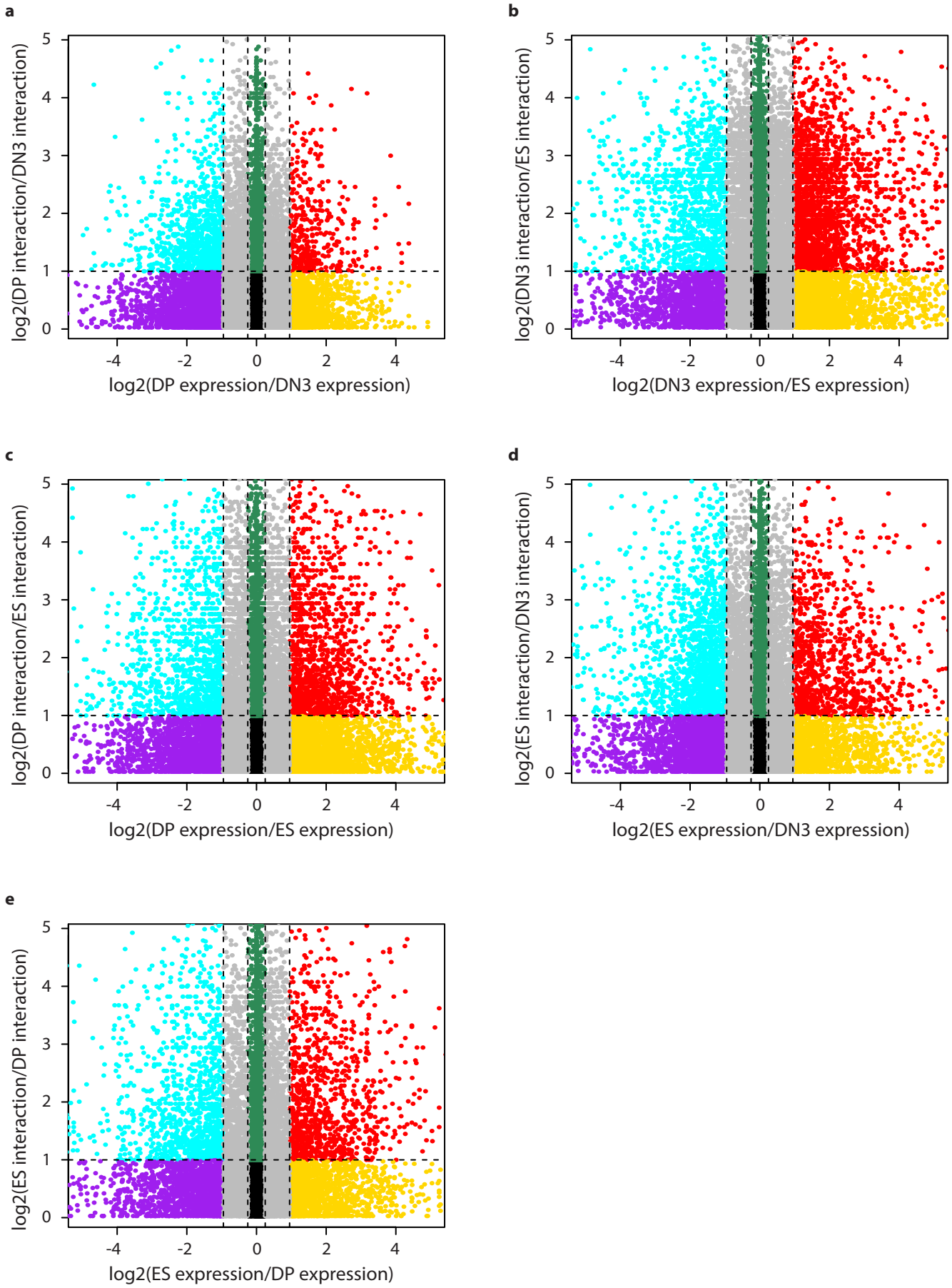


Fig S4. Stable and dynamic promoter interactions linked to transcriptional activation and repression. Scatter plots for all interactions called with b) DN3 promoters, a,c) DP promoters, or d,e) ES promoters, plotting pairwise differences in CHi-C interaction score against differences in gene expression, as computed from RNA-seq results. Different classes of interactions are labeled in different colors: A (red), B (gold), C (cyan), D (purple), E (black), F (dark green).

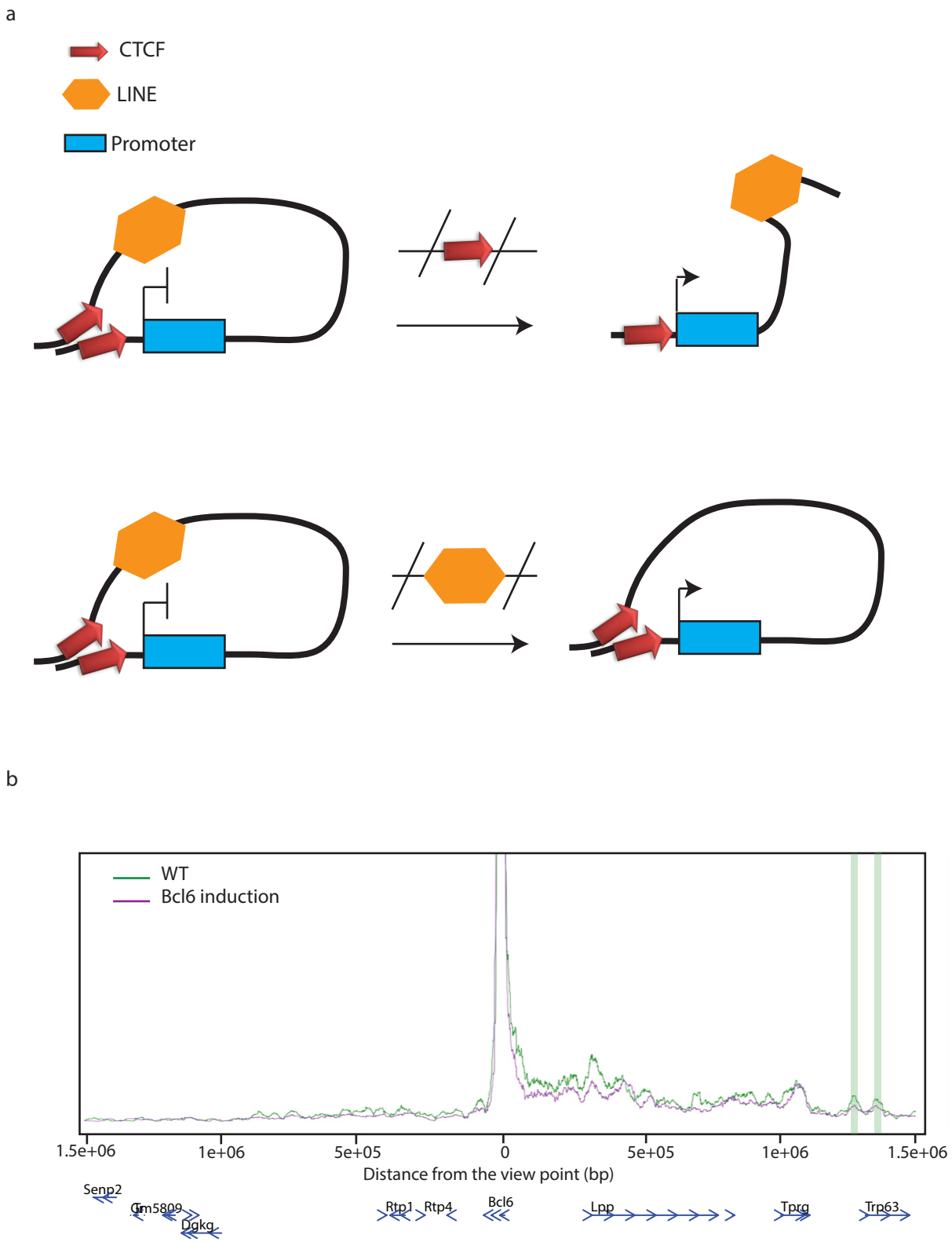


Fig S5. Following up the link between LINEs and putative silencers.

a) Schematic of ongoing CRISPR deletion experiments, addressing working hypothesis that a distal LINE (orange hexagon) next to a facing CTCF site (red arrow) is able to confer transcriptional inhibition at a contacted gene promoter (blue rectangle). Top: deletion of the distal CTCF may perturb chromatin interaction, causing derepression of gene which is no longer brought close to the LINE. Bottom: Deletion of the LINE may not affect the CTCF- mediated chromatin loop, but still cause derepression of the contacted gene. b) Preliminary 4C-seq profile of Bcl6 in ES cells before (green) and after (purple) ectopic induction. Putative promoter-silencer interactions (indicated by green stripes; these regions have been validated as silencers in luciferase reporter assays) appear to be reduced on Bcl6 induction.

Figure Legends

Fig 1. The mouse thymocyte promoter interactome.

a) Venn diagram of called promoter interactions in DN3 (blue), DP (red) and ES (green) cells. **b)** Distribution of numbers of DN3 promoter-centered interactions, classed by gene expression of target gene. **c**) Distribution of DN3 promoter-centered interaction distances, classed by gene expression of target gene. **d**) Proportions of DN3 promoter-centered interactions containing peaks for different histone modifications or bound factors. **e**) Relative enrichment of DN3 (blue) and DP (red) promoter-interacting regions for various histone marks or bound factors. **f**) CHi-C profile (DN3 blue, DP red) for local interactions with *Ikzf1* gene. **g**) DN3 (blue) and DP (red) H3K27ac ChIP-seq profile for the same genomic region. **h**) 4C profile for interactions with the *Ikzf1* promoter in DN3 (blue) and DP (red) cells. A called interaction between *Ikzf1* and a putative enhancer in both DN3 and DP cells is denoted by a purple stripe.

Fig 2. Stable and dynamic promoter interactions linked to transcriptional activation and repression.

a) Scatter plot for all interactions called with DN3 promoters, plotting difference in CHi-C interaction score between DN3 and DP against difference in gene expression between DN3 and DP, as computed from RNA-seq results. Different classes of interactions are labeled in different colors: A (red), B (gold), C (cyan), D (purple), E (black), F (dark green). **b**) Heat map showing relative enrichment or depletion (on $\log_2$ scale) of different histone marks and bound factors in regions corresponding to DP interactions of different classes, called as in **a**, except that classes E and F are further categorized into those with active (E_a, F_a) or silent (E_s, F_s) genes. **c-h**) CHi-C screenshots (DN3 in blue; DP in red) for DN3 interactions of different classes: **c**) A with *Runx1*; **d**) B with *Hes1*; **e**) C with *Dek*; **f**) D with *Cd8b1*; **g**) E with *Pten*; **h**) F with *Pik3r1*. ChIP-seq profiles for H3K27ac or H3K4me2 are shown alongside (DN3 in

blue; DP in red). Different colored stripes indicate different interactions (blue DN3-specific; red DP-specific; purple conserved in both thymocytes).

Fig 3. Thymocyte-specific and dynamic enhancers.

a) Proportion of total strong (black) and weak (white) P5424 STARR-seq hits (from Vanhille et al., 2015) present in DN3 promoter interactions, classed according to interaction type. **b**) Luciferase reporter assay results in P5424 (purple) and ES cells (green), expressed as fold increase in reporter expression over minimal reporter constructs. Results are shown for a constitutive (SV40), and ES-specific (Sox2) controls, as well as CHi-C-called thymocyte-specific and DN3-specific interacting enhancers. *** $P < 0.001$; two-tailed t-test comparing the two cell types, with Benjamini-Hochberg multiple testing correction. **c**) CHi-C profile (blue DN3, red DP) around the *Myc* gene. **d**) H3K27ac ChIP-seq profiles (blue DN3, red DP) around the same region. **e**) 4C profile for interactions with the *Myc* promoter (blue DN3, red DP). **f**) qRT-PCR results for *Myc* expression, expressed relative to actin, in P5424 cells after treatment with dCas9-KRAB and guide RNAs directed to either the *Myc* enhancer or an unrelated genomic region as control.

Fig 4. Distal silencers may regulate contacted genes.

a) Luciferase reporter assays for 2 kb test inserts upstream of the SV40 enhancer/promoter in plasmids transfected in P5424 (purple) or ES (green) cells. Reporter expression is expressed as proportion of the SV40 enhancer/promoter construct without other insert. *P*-values are calculated by two-tailed t-tests comparing the test insert with its corresponding cell type-matched neutral sequence, with Benjamini Hochberg multiple testing correction. *** $P < 0.001$; ** $P < 0.005$; * $P < 0.05$ **b**) Proportion of interacting regions containing a particular class of TE adjacent to a CTCF motif, classed according to expression of the interacting gene. *P*-values are given from the Fisher's exact test, comparing the silent gene interactions with all active gene interactions. **c**) Luciferase reporter assays for ~500 bp test inserts upstream of the

SV40 enhancer/promoter in plasmids transfected in P5424 (purple) or ES (green) cells, as in **a**). The TEs present within these test regions are denoted.

Supplemental Data

Supplemental Figures

Fig S1. The mouse thymocyte and ES promoter interactome features.

a) Distribution of numbers of DP promoter-centered interactions, classed by gene expression of target gene. **b**) Distribution of DP promoter-centered interaction distances, classed by gene expression of target gene. **c**) Distribution of numbers of ES promoter-centered interactions, classed by gene expression of target gene. **d**) Distribution of ES promoter-centered interaction distances, classed by gene expression of target gene. **e**) Proportions of DP promoter-centered interactions containing peaks for different histone modifications or bound factors. **f**) Relative enrichment of DP promoter-interacting regions for selected histone marks or bound factors. **g**) Total percentage and **h**) relative enrichment of different cell type promoter interactions, and interactions conserved in all cell types, for conserved CTCF sites.

Fig S2. Stable thymocyte promoter interactions.

a-e) Selected CHi-C and corresponding 4C profiles for interactions which are predominantly conserved in DN3 and DP cells. Called interactions are given by purple (conserved), red (DP) or blue (DN3) stripes.

Fig S3. Dynamic thymocyte promoter interactions.

a-d) Selected CHi-C and corresponding 4C profiles for interactions which are cell type-specific. Called interactions are given by red (DP) or blue (DN3) stripes.

Fig S4. Stable and dynamic promoter interactions linked to transcriptional activation and repression.

Scatter plots for all interactions called with **b**) DN3 promoters, **a,c**) DP promoters, or **d,e**) ES promoters, plotting pairwise differences in CHi-C interaction score against differences in gene

expression, as computed from RNA-seq results. Different classes of interactions are labeled in different colors: A (red), B (gold), C (cyan), D (purple), E (black), F (dark green).

Fig S5. Following up the link between LINEs and putative silencers.

a) Schematic of ongoing CRISPR deletion experiments, addressing working hypothesis that a distal LINE (orange hexagon) next to a facing CTCF site (red arrow) is able to confer transcriptional inhibition at a contacted gene promoter (blue rectangle). Top: deletion of the distal CTCF may perturb chromatin interaction, causing derepression of gene which is no longer brought close to the LINE. Bottom: Deletion of the LINE may not affect the CTCF-mediated chromatin loop, but still cause derepression of the contacted gene. **b**) Preliminary 4C-seq profile of *Bcl6* in ES cells before (green) and after (purple) ectopic induction. Putative promoter-silencer interactions (indicated by green stripes; these regions have been validated as silencers in luciferase reporter assays) appear to be reduced on *Bcl6* induction.

Supplemental Tables

Table S1. CHi-C interactions with DN3, DP and ES promoters.

10 first lines of Interaction file of DN3

ID bait	Chr	Start	End	Gene	Chr	Start	End	ID OE1	ID OE2	Reads rep1	Reads Rep2
230135	chr5	31347443	31351247	0610007C21Rik	chr5	31337369	31338974	230133	230133	223	200
260822	chr5	130691646	130696119	0610007L01Rik	chr5	130685715	130689249	260820	260820	71	61.5
571305	chr13	63915555	63920989	0610007P08Rik	chr13	63182813	63215504	571054	571066	14	10.5
571305	chr13	63915555	63920989	0610007P08Rik	chr13	63494501	63517485	571168	571157	11	8.5
571305	chr13	63915555	63920989	0610007P08Rik	chr13	64475897	64488020	571502	571508	6	4.5
115212	chr2	163362868	163370456	0610008F07Rik	chr2	163374689	163380126	115217	115216	76	74.5
115212	chr2	163362868	163370456	0610008F07Rik	chr2	163536784	163540059	115264	115264	24	19.75
493425	chr11	51500281	51503076	0610009B22Rik	chr11	50212059	50236964	493043	493036	9	7.5
493425	chr11	51500281	51503076	0610009B22Rik	chr11	51109324	51110246	493323	493323	16	13.5

10 first lines of Interaction file of DP

ID bait	Chr	Start	End	Gene	Chr	Start	End	ID OE1	ID OE2	Reads rep1	Reads Rep2
230135	chr5	31347443	31351247	0610007C21Rik	chr5	31337369	31338974	230133	230133	177	200
260822	chr5	130691646	130696119	0610007L01Rik	chr5	130685715	130689249	260820	260820	58	65
571305	chr13	63915555	63920989	0610007P08Rik	chr13	63655962	63663019	571222	571219	15	23.5
514148	chr12	4823453	4824477	0610009D07Rik	chr12	3783618	3790397	513840	513839	12	15.5
514148	chr12	4823453	4824477	0610009D07Rik	chr12	4824727	4827786	514150	514150	130	138.5
498904	chr11	70049372	70051622	0610010K14Rik	chr11	68890022	68911166	498645	498638	15	20
498904	chr11	70049372	70051622	0610010K14Rik	chr11	69819574	69851418	498838	498846	24	35.5

498904	chr11	70049372	70051622	0610010K14Rik	chr11	70047137	70048029	498902	498902	187	206.5
700308	chr17	26009741	26014230	0610011F06Rik	chr17	25986736	25996602	700303	700303	79	97
113123	chr2	156372010	156373689	0610011L14Rik	chr2	156538227	156543195	113160	113160	26	36.25

10 first lines of Interaction file of mESCs

ID bait	Chr	Start	End	Gene	Chr	Start	End	ID OE1	ID OE2	Reads rep1	Reads Rep2
230135	chr5	31347443	31351247	0610007C21Rik	chr5	31337369	31338974	230133	230133	165	192
260822	chr5	130691646	130696119	0610007L01Rik	chr5	130685715	130689249	260820	260820	116	94
571305	chr13	63915555	63920989	0610007P08Rik	chr13	63607546	63617838	571203	571202	5	24.5
571305	chr13	63915555	63920989	0610007P08Rik	chr13	64168909	64210194	571396	571395	4	17.5
540967	chr12	87163534	87166516	0610007P14Rik	chr12	87156936	87162461	540965	540965	134	153.5
540967	chr12	87163534	87166516	0610007P14Rik	chr12	87415220	87433849	541048	541052	11	21.5
115212	chr2	163362868	163370456	0610008F07Rik	chr2	163072583	163075978	115137	115137	14	19.75
484158	chr11	23530271	23535902	0610010F05Rik	chr11	23782795	23795635	484250	484252	21	16
733115	chr18	36503708	36506376	0610010O12Rik	chr18	36271470	36272747	733028	733028	16	22.5
733115	chr18	36503708	36506376	0610010O12Rik	chr18	36385748	36388904	733077	733077	40	54.5

Table S2. Source of epigenomic datasets used in this analysis.

	DP	DN3	mESc
H3k4me1	GSM523698	GSM756894	GSM1359829
H3K4me3	GSM523699	GSM1872304	GSM723017
H3K27ac	GSM1556287	GSM2113441	GSM851278
H3K27me3	GSM1818900	GSM1498422	**GSM1000089**
RNAseq	GSM727007; GSM727007	GSM1649842; GSM1649849	**GSM723776**
PolII	GSM726991	GSM1340641	GSM723019
Ikaros	GSM1498444	GSM1498442	No data
CTCF	GSM672400	GSM1023416	GSM723015
ETS1	GSM726992	GSM1360719	Not data
H3K122ac	No data	No data	GSE66023
H3K64ac	No data	No data	GSE66023
RunX1	GSM1095815	GSM1360735	No data
Cohesion	**GSM1184316**		
H3K9me3		Total thymocytes **GSM945744**	

IV. TADs caller benchmarking

In order to identify the appropriate tool for TAD calling in CHi-C data, I benchmarked different tools. A comparative analysis suggests the use of Arrowhead algorithm to call TADs in CHi-C (TADs).

1. TAD calling tools

1.1 TADbit

TADbit (alpha version 360) uses a breakpoint detection algorithm which is commonly used for detection of copy-number variants (Serra, Baù, Filion, & Marti-Renom, 2016). It identifies the optimal segmentation of chromosome into domains under a Bayesian information criterion (BIC) penalized likelihood. It is a python package with different tools for read alignments, normalization, TAD identification and compartment calling. In this study, we only use its TAD calling tool. As input, TADbit requires a symmetric matrix of observed counts which are automatically normalized using a modified version of ICE (Imakaev et al., 2012) normalization called “Visibility normalization”.

For TAD-Capture analysis, we transformed each normalized matrix in the ibed format (see previous section on TAD-capture normalization) into a symmetric matrix (each line and column corresponds to unique bin: “chromosome Number_ID of the bin”), using a custom perl script. Two parameters are important for TADbit, the maximum TAD size (default is the entire chromosome length) and the possibility to identify centromeric regions. Here, we kept default TAD size and we set the parameter to identify centromeric regions to TRUE.

1.2 The Insulation score

TADs are demarcated by boundaries which are known to be enriched in insulator binding.
Thus, for each genomic position in a given resolution, a boundary is defined as the genomic

region with high insulation strength. Based on this idea, the insulation score (v1.0.0) (Crane et al., 2015) calculates an insulation score for every genomic region by using a sliding window of contact signals along the diagonal. Therefore each bin along the diagonal is assigned with an insulation score. Based on the insulation vector, an insulation delta vector is further calculated using as second sliding window which shows the difference between the left and right of each bin. A TAD boundary is then defined as the bin containing the local maximum of insulation delta score.

For TAD-Capture data, the insulation square was set to 50 kb for our datasets with 5 kb resolution. The insulation delta span was set to 20 kb. Default settings were used for insulation mode, noise 16 threshold and boundary margin of error (mean =0.1). The output is the insulation score and the delta values for each bin plus the coordinates of called boundaries of whole genomic region except the first and the last portion of the matrix which corresponds to the size of insulation square.

1.3 Armatus

Armatus (Filippova, Patro, Duggal, & Kingsford, 2014b) (v2.0) is based on a multiscale approach which identifies a set of consensus domains across different resolutions. It uses a score function that calculates the local density of intra domain interactions at different resolutions defined by the user (gamma parameter). Depending on the calculated score, the algorithm finds a consensus set of TADs that persists across various resolutions.

Armatus is implemented in C++ language and requires a complete preprocessing pipeline to generate the normalized matrix.

To call TADs on TAD capture matrices, first, we transformed the ibed format of TAD-capture matrices into the appropriate input of Armatus which is a symmetric matrix of observed or normalized entries. We set gamma-max to 0.05 and we kept all other parameters to their default setting. We ran the tool on local server and it took 2 mins to call TAD borders for each

matrix. The output is a file of 3 columns with each line representing a consensus TAD: The first column is the chromosome number and the next two columns are the start and ending indices of bins in a domain.

1.4 Arrowhead

Arrowhead is one part of other tools in the juicer pipeline used for Hi-C data analysis (Durand et al., 2016). It is based on the arrowhead transformation algorithm which transform squares along the diagonal of Hi-C contact map into triangles. The idea behind is that squares are a complex and difficult shape to detect while triangles are easier to identify. The transformation results in arrows-like patterns of High and low signal for each square (TAD) along the diagonal. The algorithm, then, computes specific score "corner score" (based on: the sign of triangles, the sum of entries and the variance of entries) for the triangles designed around the pair of loci to assess their potential as TAD boundaries. Therefore, TADs are determined at different level of hierarchy by using dynamic programming algorithm.

Arrowhead takes as input the .hic file produced by Juicer Tools Pre. Here, we converted the normalized matrices into .hic files using Juicer Tools Pre imposing no normalization (-n parameter). To call TAD borders, arrowhead was used with default parameters except the normalization (-K = set to NONE) and the resolution parameter (-r set to 5 kb).

2. Arrowhead for TADs calling in CHi-C (TADs)

To compare TAD callers on experimental data (CHi-C TADs), I considered the total number of called TADs, the TAD size and the visual concordance of identified TADs (**Fig 1**). The number of TADs identified varied from tool to tool (**Fig 1B**). On average, in all data sets at 5 kb resolution, Armatus (Filippova, Patro, Duggal, & Kingsford, 2014a) called the largest (180) and Insulation score (Crane et al., 2015) the smallest (36) number of TADs. Noting that TADbit (Serra et al., 2016) and Insulation score partition chromosomes in a continuous set of TADs, whereas the others allow gaps between TADs like Arrowhead (Durand et al., 2016)

which adopt multiscale approaches returning nested TADs. Thus, Armatus (Filippova et al., 2014a) returned TADs with small size whereas Arrowhead (Durand et al., 2016) returned the biggest TADs (**Fig 1B**). By visual inspection, Arrowhead (Durand et al., 2016) seems the only tool that detected almost all TADs present in CHi-C data **(Fig 1A**). This is due to only dependence of Arrowhead on the coverage of the interaction matrix, very high coverage in CHi-C matrices. Whereas, other tools mostly require a window size set to identify TAD borders. Although, a recent study (Forcato et al., 2017) comparing between different TADs caller in Hi-C, suggests no single method outperforms others in all situations, Arrowhead (Durand et al., 2016) visually outperforms other tools in CHi-C (TADs).

Finally, a robust quantification of performance in terms of specificity and sensitivity is hindered by the lack of ground-truth-positive and ground-truth-negative controls for chromatin architecture and by conceptual difficulties in designing simulators of Hi-C data.

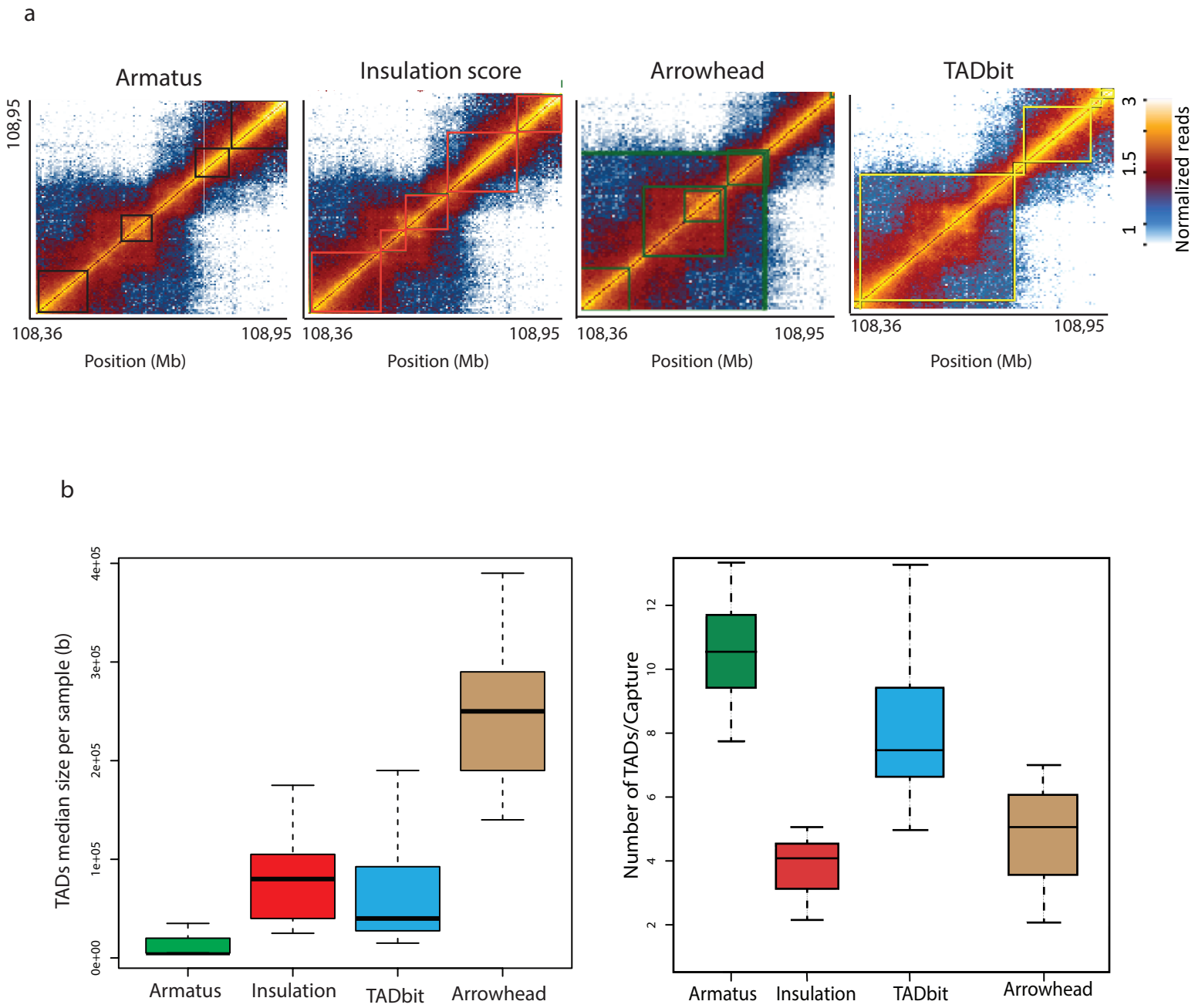


Fig 1: Comparative results of methods for the identification of TADs.
a) Heatmap of the contact matrix of DN3 (chr8:108,360,000-108,950,000) at 5kb resolution. Identified TADs are framed in different colors for the various methods.
b) Boxplot of median TAD size in all replicates of all datasets (analyzed at 5kb). Boxplot of number of TADs per capture region in all replicates of all datasets (analyzed at 5kb).

Transcription directly remodels a small subset of topologically associated domains

Sanjay Chahar[1-5], Yousra Ben Zouari[1-5], Anne M Molitor[1-4], Dominique Kobi[1-4], Manon Maroquenne[1-4], Audrey Mossler[1-4], Nezih Karasu[1-4] & Tom Sexton[1-4]*

[1] Institute of Genetics and Molecular and Cellular Biology (IGBMC), Illkirch, France

[2] CNRS UMR7104, Illkirch, France

[3] INSERM U1258, Illkirch, France

[4] University of Strasbourg, Illkirch, France

[5] These authors contributed equally to this work.

* Correspondence should be addressed to T.S. (sexton@igbmc.fr).

Preface

This work was a large collaborative effort within the group. I performed all of the computational analysis of the CHi-C data and comparison with all other genome-wide datasets.

Abstract and Introduction

Metazoan genomes are spatially organized into self-folded topological associated domains (TADs), which have been proposed to demarcate functional genomic units, based on correlation with epigenomic profiles (Dixon et al., 2012; Sexton et al., 2012) and the apparent constraints they place on the operational range of enhancers (Symmons et al., 2014; Lupianez et al., 2015; Symmons et al., 2016). A physical model comprising loop extrusion by cohesin, with CTCF-bound sites defining TAD borders as barriers to the extrusion, explains much of the observations of Hi-C datasets (Fudenberg et al., 2016; Sanborn et al., 2016), and is supported by elegant studies perturbing cohesin and/or CTCF (Haarhuis et al., 2017; Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017), providing a general mechanism for TAD creation and maintenance. However, despite our growing appreciation that TADs can regulate gene expression in physiological and pathological situations (Le Dily et al., 2014; Lupianez et al, 2015; Franke et al., 2016), we have little understanding as to if or how loop extrusion is modulated to accommodate or influence transcriptional changes of the underlying genes. Initial comparative Hi-C studies concluded that TADs were largely tissue-invariant (Dixon et al., 2012; Dixon et al., 2015), suggesting that they are stable architectures on which finer-scale transcriptional regulation is overlaid. However, higher-resolution views have identified counter-examples of TADs or "sub-TADs" which are remodeled in line with transcriptional changes in the underlying genes (Noordermeer et al., 2011; Phillips-Cremins et al., 2013; Bonev et al., 2017). Since TAD borders are enriched in active genes (Dixon et al., 2012; Sexton et al., 2012), it has been proposed that the local topological changes brought about by RNA polymerase binding and elongation (Lavelle, 2014) could impact on higher-order folding into TADs, presumably by local chromatin decondensation and/or modulating barriers to loop extrusion. However, the causal role of transcription in TAD creation or maintenance remains disputed. TAD appearance in early embryogenesis correlates with

zygotic genome activation, but was unaffected by transcriptional inhibition (Du et al., 2017; Hug et al., 2017; Ke et al., 2017). Further, ectopic induction of a gene in mouse embryonic stem (ES) cells failed to recapitulate the TAD remodeling which was observed to accompany this gene's activation during neuronal differentiation (Bonev et al., 2017).

To analyze TAD architecture at high resolution during a developmental transition, we used oligonucleotide capture coupled to *in situ* Hi-C (CHi-C) to interrogate the local chromatin architecture around several genes which are highly up- or downregulated during mouse thymocyte maturation. We found that the majority of interrogated TADs and sub-domains were unchanged in these different cell types, even around genes with over 6-fold increases in expression. Nevertheless, a subset of domains were remodeled concomitantly with gene activation, either by the shift of a boundary to accommodate the fully transcribed gene, or the creation of a new sub-domain comprising the active gene unit. In the latter case, ectopic induction was sufficient to drive partial TAD remodeling. This provides the first evidence, to our knowledge, of direct TAD control by transcription, suggesting that gene expression is one of likely many mechanisms regulating chromatin architecture.

Results

Predominant TAD conservation during thymocyte maturation

We performed CHi-C in mouse $CD4^-$ $CD8^-$ $CD44^-$ $CD25^+$ (double negative; DN3) and $CD4^+$ $CD8^+$ (double positive; DP) thymocytes, using a capture strategy of tiled oligonucleotides covering nearly all the restriction fragments (with the four-cutter DpnII) within eight ~600 kb regions spanning genes of interest located very close (<20 kb) to called TAD borders in mouse ES cells (Dixon et al., 2012) (**Table 1**). These cell types represent populations just before and after the checkpoint for productive rearrangement of the T cell receptor-β gene, which is essential for generating productive T cells and is accompanied by well-characterized

transcriptional and epigenomic changes at hundreds of gene loci (Egawa and Littman, 2011; Koch et al., 2011; Pekowska et al., 2011; Zhang et al., 2012). Three captured regions are centered on genes (*Bcl6*, *Nfatc3*, *Rag1*) that are upregulated, three on genes (*Cdh1*, *Il17rb*, *Pla2g4a*) that are downregulated, and two on genes (*Cd3g*, *Zap70*) which have unaltered expression on the DN3-to-DP transition. For comparison with an unrelated cell type, we also performed the same CHi-C in mouse ES cells. As expected, for comparable sequencing depths, CHi-C gave much higher coverage and resolution at the interrogated regions than conventional Hi-C, allowing some chromatin loop interactions to be distinguished and a higher-confidence calling of TAD borders (**Fig S1**). By both visual inspection and computational calling of TAD borders with the arrowhead algorithm (Rao et al., 2014), TAD architectures were largely unchanged in all three cell types examined, regardless of clear large transcriptional differences at the genes within some of these regions (**Fig 1a**). When comparing the numbers of TAD borders that are exactly identical (at 5 kb resolution) across the cell types, just over half appeared unique to one particular cell type (**Fig 1b**). However, visual inspection suggests that many of these are actually conserved, but that the TAD border calling algorithm can vary by one or two pixels, both when comparing cell types or the very reproducible biological replicates. As a result, using the exact intersection likely underestimates the true number of conserved TAD borders. Rather than trust an arbitrary threshold of pixel proximity for whether a TAD border is conserved or not, we are currently exploring this problem in more detail, benchmarking the thresholds against the biological replicates. The few cell type-specific changes in TAD borders that were readily identified on visual inspection were reproducible across the highly consistent biological replicates (e.g. **Fig S2**). Interestingly, “stable” TAD borders were much more highly enriched in CTCF binding than more tissue-specific borders (**Fig 1c**), supporting the protein’s role as an “architectural” protein (Phillips and Corces, 2009).

Transcription can drive sub-TADs around gene units

The highest-resolution genome-wide appraisal of developmental chromosome folding dynamics to date identified a number of cell type-specific TAD boundaries at the transcription start sites (TSS) of upregulated genes (Bonev et al., 2017). Within the interrogated regions, we also observed two cell type-specific TAD borders at the TSS of differentially expressed genes: one at the promoter of *Nfatc3*, highly expressed in DP cells, and one at the promoter of *Tmem131*, which has much higher expression in DN3 cells (**Figs 2,3**). In both cases, the new border was observed in cells where the underlying gene was most active. Direct comparison of the normalized CHi-C contact strengths across the two cell types reveals increased intragenic interactions across the whole gene body on transcriptional activation, suggesting that the gene forms a topological sub-domain, rather than just the TSS acting as an isolated barrier or "insulator". Although active gene units have been suggested to form spatial domains in yeast (Hsieh et al., 2015) and metazoans (Rowley et al., 2017), their genome-wide prevalence has not been supported in the majority of high-resolution Hi-C studies (Rao et al, 2014; Bonev et al., 2017). Indeed, for the two thymocyte subtype-specific domains we identified by CHi-C, many other differentially expressed genes had no measurable changes in chromatin topology (**Fig 1**), suggesting that transcriptional induction is rarely sufficient to remodel TADs. Curiously, the 3' ends of *Nfatc3* and *Tmem131* form TAD borders that are conserved in ES, DN3 and DP cells, and only the TSS forms a developmentally dynamic border. It is thus possible that gene induction can only efficiently remodel topological domains at regions where the architecture is already pre-disposed by other mechanisms.

As mentioned previously, a direct role of transcription in defining TADs is hotly debated (Du et al., 2017; Hug et al., 2017; Ke et al., 2017). In a more direct test, ectopic induction of two genes, *Zfp608* and *Sox4*, whose TSSs were observed to form new TAD borders when the genes were activated on neural differentiation, failed to alter chromatin

architecture (Bonev et al., 2017). However, these two genes differed from those coming out of our CHi-C studies, in that the topological changes on differentiation did not encompass the whole gene body and appeared restricted to the TSS acting as an "insulator". To determine whether transcription can directly remodel TADs in other gene contexts, we used nuclease-dead Cas9 fused to the transcriptional activation domain VP64 (Konermann et al., 2015) to target the ectopic induction of *Nfatc3* in ES cells, where the gene is silent and does not form a spatial domain. Targeting the *Nfatc3* promoter with four guide RNAs induced an almost 6-fold increase in gene expression and, importantly, was sufficient to create a topological sub-domain comprising the gene body (**Fig 4**). Direct comparison of normalized CHi-C contact maps showed that the position of the new domain on ES induction is identical to that arising in the DN3-to-DP transition, although it is quantitatively weaker, suggesting that even in this case, other mechanisms are required to reinforce chromatin topology. We have performed similar experiments for *Tmem131* induction in ES cells, and CHi-C experiments are ongoing. This is the first evidence, to our knowledge, that TADs can be directly remodeled by transcription, albeit in very specific genomic contexts.

TAD borders can shift to accommodate transcriptional regulatory events

We also observed an interesting chromatin topological change at the *Bcl6* gene (**Fig 5**). Although initial microarray studies classed this gene as essentially silent in DN3 cells (Egawa and Littman, 2011), chromatin immunoprecipitation studies actually revealed a large amount of paused RNA polymerase at the promoter-proximal region. Interestingly, this RNA polymerase peak corresponds to a cluster of CTCF sites and a clear TAD border in DN3 cells (**Fig S3**). When *Bcl6* is fully activated in DP cells, the TAD border relocates by ~20 kb to beyond the 3' end of the gene, allowing a single domain to now encompass the whole transcribed unit. As well as full gene transcription, this topological change is concomitant

with the appearance of active histone marks at a putative upstream enhancer, which forms DP-specific looping contacts with the *Bcl6* promoter (**Fig 5**). Although they are more transient and thus harder to catch by 3C methods, enhancers have been reported to make contacts with gene bodies as well as promoters, perhaps somehow linked to the tracking of engaged RNA polymerase (Lee et al., 2015). The observed TAD border shift could thus be caused directly by RNA polymerase elongation, analogous to the sub-domain created at induced *Nfatc3*, or be the result of accommodation of the enhancer into the active chromatin hub. We used the dCas9-VP64 system to ectopically induce *Bcl6* in ES cells where the gene is completely silent, with no paused polymerase or upstream enhancer, and the TAD border is identical to DN3 cells (**Fig S4**). Despite a more than 30-fold induction of *Bcl6*, the border was completely unchanged, suggesting that in this genomic context, transcription is insufficient for TAD remodeling, and that perhaps the upstream enhancer interactions play a more important architectural role.

We next asked what mechanisms other than transcription could be responsible for the exact location of the TAD borders around *Bcl6* at the DN3-to-DP transition. Interestingly, comparison of quantile-normalized CTCF ChIP-seq datasets for DN3 and DP cells (Shih et al., 2012) revealed the presence of CTCF binding at both the DN3 and DP TAD borders, with an apparent quantitative change in CTCF binding preference according to cell type, concordant with the choice of TAD border (**Fig S3**). In ES cells, CTCF is readily found at the *Bcl6* promoter, but not at the downstream site. We hypothesize that the CTCF site downstream of the *Bcl6* gene is a "secondary" TAD border that is employed when the principal one is made unavailable by the transcriptional processes occurring at *Bcl6* in DP cells. To test this, we have made ES cells with a homozygous deletion of the CTCF motifs at the "primary" site and are testing by CHi-C, ChIP-qPCR and qRT-PCR whether there is any effect on chromatin topology, CTCF binding to the secondary site, and/or *Bcl6* expression,

respectively. We are also testing whether ectopic *Bcl6* induction affects CTCF binding at either of the sites, and are re-assessing the quantitative CTCF binding differences in DN3 and DP cells by ChIP-qPCR.

Towards a genome-wide assessment of developmental TAD dynamics

Although CHi-C offers an unparalleled resolution of chromatin interactions for a given sequencing depth (**Fig S1**), our approach is limited to a handful of TADs. From these, we identified three interesting cases of transcriptional regulation-linked TAD remodeling during thymocyte maturation, but do not know if these are the *only* examples, or whether these observed phenomena represent a significant minority of developmental TADs genome-wide. For each CHi-C experiment, we sequenced a corresponding pre-capture Hi-C sample, initially for quality control purposes (see earlier chapter in Results). We also performed a similar strategy for promoter CHi-C in DN3 and DP cells (see earlier chapters in Results), although this time with the six-cutter enzyme HindIII. When pooling all Hi-C datasets for a particular thymocyte subset together, we obtained interaction maps of sufficient coverage that the previously described remodeling events at *Bcl6*, *Nfatc3* and *Tmem131* could also be observed, at an approximate resolution of 20 kb (**Fig S5**). We initially performed the arrowhead algorithm (Rao at el., 2014) on these pooled Hi-C datasets to call their TAD borders in an identical manner to that for the CHi-C experiments. However, the numbers of TADs robustly called by this method for the sparser Hi-C maps was far fewer than we observed on visual inspection. We are currently carefully benchmarking other TAD calling methods (see also Chapter) to try and come up with the most reliable list possible of TADs which are remodeled during thymocyte maturation. We will then see to what extent our hypothesized factors determining the more easily remodeled TADs (for example, the presence of a pre-formed border at the 3' end of the gene) are applicable genome-wide. This approach will also be

performed in parallel on Hi-C maps charting neuronal differentiation of ES cells, to see if such observations also hold in other differentiation models.

Discussion

Our CHi-C strategy, focusing on specific TADs during thymocyte maturation, has confirmed findings from initial, lower-resolution studies that the majority of topological domains appear invariant to expression changes at their underlying genes (Dixon et al., 2015). However, like other recent studies (e.g. Bonev et al., 2017), we observed specific genomic contexts where transcriptional induction or upregulation is correlated with spatial chromosomal remodeling. We observed two different behaviors: the generation of sub-domains corresponding to entire transcribed gene units; and the shifting of TAD borders from a location where a gene is split between two domains, to one where the transcribed gene is contained within a single TAD. For the former class, we provide the first direct evidence to date that transcriptional induction is causal in TAD restructuring, at least in specific genomic contexts. A major question is why, despite the large disruption of nucleosome structure that presumably accompanies processive elongation of RNA polymerase (Lavelle, 2014), the majority of TAD architectures appear refractory to underlying gene expression changes. For many genes, transcriptional firing may be a sufficiently rare event, and/or the gene is too short for any small or brief topological disruptions to be resolved by population-average (C)Hi-C approaches, but this is unlikely to explain all observed stable TADs. Greater mechanistic appraisal and understanding of the cohesin loop extrusion model is required to better predict if and how elongating RNA polymerase can modulate or interfere with cohesin loading, unloading or extrusion.

Although not fully supported by other Hi-C studies, it has been proposed that active gene units can make up individual small topological domains (Hsieh et al., 2015), and that these could even represent the well-described developmental shifts of genomic regions

between the cross-interacting active ("A") and inactive ("B") compartments at a finer scale (Rowley et al., 2017). Due to the paucity of active genes that seem to readily form such new domains, it appears likely that such effects are relatively weak and easily overridden by more direct architectural principles, such as loop extrusion. Hopefully a better genome-wide view of which active genes can form new domains will provide better clues as to what mechanistically distinguishes invariant from more malleable TADs.

The other type of TAD dynamics observed at the *Bcl6* gene is intriguing, and raises the question of what function TAD borders placed inside genes may play. The *Bcl6* upstream enhancer does not carry active histone marks until the DP stage, so this border is not likely to be necessary to prevent aberrant promoter-enhancer communication, as posited for other TAD borders (Lupianez et al., 2015). Further, the intragenic border is conserved in ES and DN3 cells, even though one cell type is able to completely silence the gene and the other is able to accumulate paused polymerase at the promoter. Border perturbation experiments are likely required in developing thymocytes to determine what, if any, role this TAD border plays on fine-tuning *Bcl6* regulation. To date, the only other description of a potential function of plastic TAD borders directly at genes has been at specific Hox genes, whereby the regulatory elements from one flanking TAD or the other are employed according to developmental timing (Andrey et al., 2013). However, the resolution of this study was insufficient to determine whether the TAD border was ever contained inside the Hox gene, or whether the whole gene swapped TAD occupancy. In any case, more detailed genome-wide views are required to assess to what extent intragenic TAD borders can be employed as a gene regulatory mechanism. It is an exciting prospect, if speculative at the moment, that topological domains do not only delimit the functional range of *cis*-regulatory elements, and/or facilitate their search for cognate genes in three-dimensional nuclear space, as has been previously proposed (Symmons et al., 2014; Sexton and Cavalli, 2015), but that the borders

themselves can also facilitate polymerase pausing, thus maintaining important genes in a poised state.

Materials and methods

Isolation of mouse DN3 and DP thymocytes

Thymuses were dissected from 6-8 week old c57/Bl6 mice, and DN3 and DP cell populations were purified by fluorescent assisted cell sorting (FACS), following the protocol of Oravecz et al. (2015).

ES cell culture and CRISPR/Cas9 genome engineering

ES cells were maintained as in Bibel et al. (2007). Deletion experiments were performed by transfecting ES cells with custom plasmids encoding Cas9 and two guide RNAs in parallel, made by the IGBMC platform. The most highly transfected cells were sorted by limited puromycin selection, followed by FACS for GFP expression. Single cells were amplified to clones and screened for deletions by PCR assays.

Hi-C and promoter CHi-C

In situ Hi-C was performed with DpnII, essentially as in Vietri Rudan et al. (2017). Capture oligonucleotides were extracted as 120-nucleotide stretches adjacent to all DpnII sites within the target regions (**Table 1**), filtering out restriction fragments that were smaller than 120 bp, or where 120 bp regions could not be found that had a mappability score greater than 90% (see Yaffe and Tanay, 2011). The CHi-C experiments with this custom oligonucleotide set (ordered as Agilent SureSelect probes) were performed essentially as Schoenfelder et al., 2015.

dCas9-VP64 induction experiments

Induction experiments were performed essentially as for Bonev et al., (2017).

Hi-C read processing and filtering

Hi-C reads were pre-processed and valid reads filtered following a pipeline very similar to that of Sexton et al. (2012). See also the previous results chapter for Hi-C quality controls.

CHi-C matrix normalization

See previous chapter for details

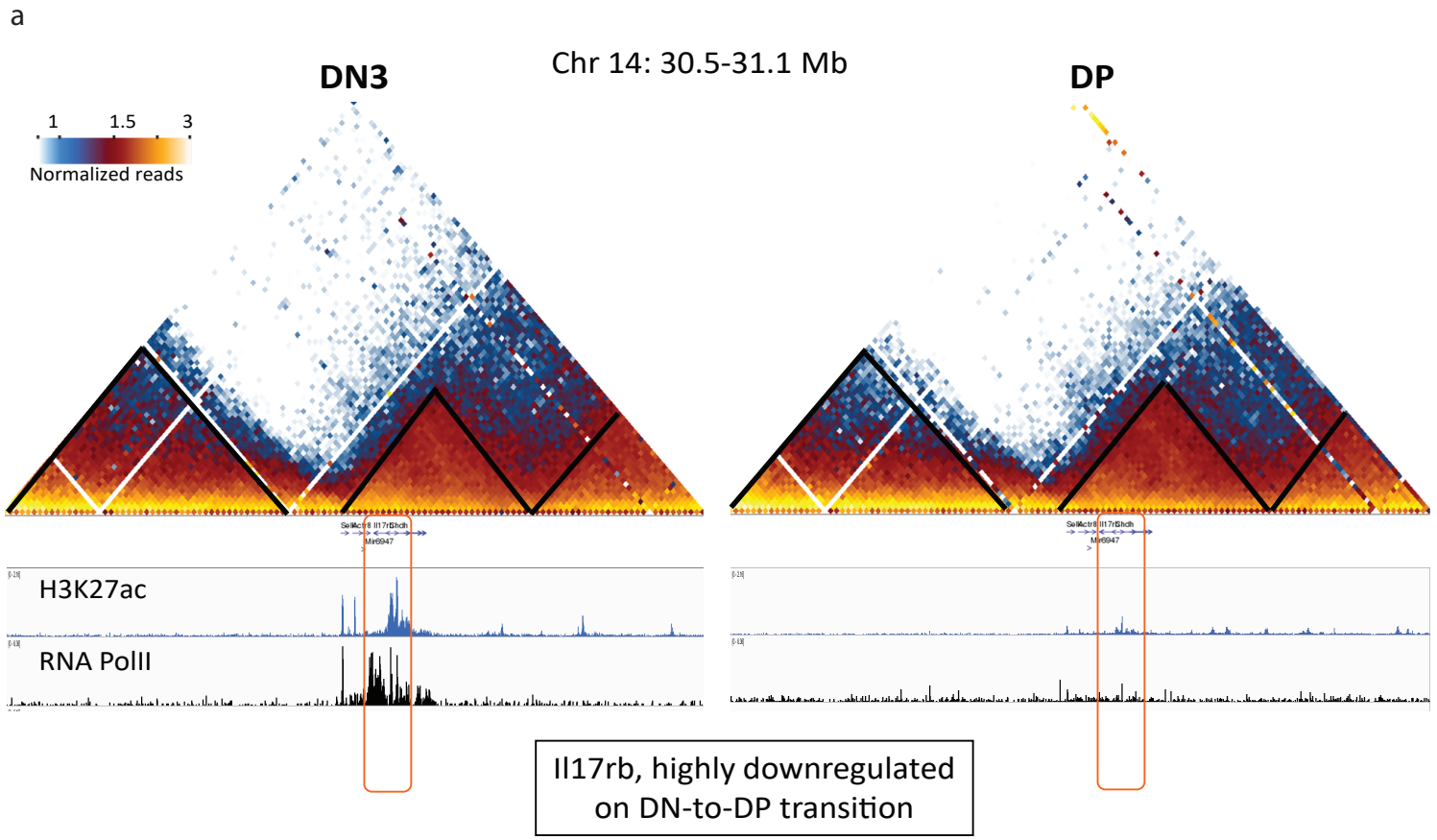
TAD calling

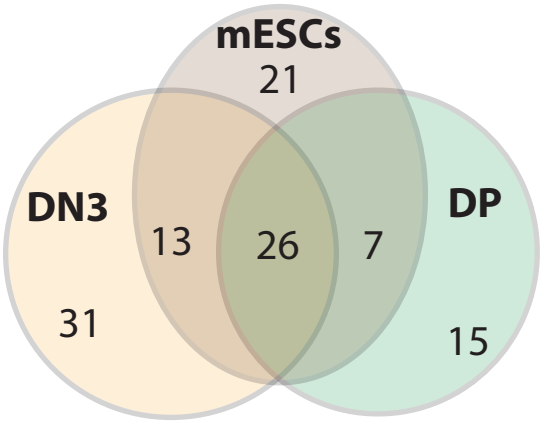
TADs were called from the CHi-C matrices by the Arrowhead algorithm (Rao et al., 2014). See previous chapter for details.

Epigenomic profile sources and pre-processing

ChIP-seq data were downloaded from GEO database for these histone marks: H3K4me1, H3K27ac, H3K4me3, H3K27me3 and for these transcription factors: Ikaros, Runx1, Ets1, GATA3 and for PolII, CTCF and cohesin. I aligned ChIP-seq on mm9 genome using bowtie2. Then, I called peaks using Erange V4.0. In Erange, mapped reads in the SAM file are first transformed into native Erange reads store .rds file. Then, peaks are identified with the peaks finder tools in Erange with –nodirectionality and –notrim parameters. Erange returns a per-peaks p-value. Wig files of Histone marks and transcription factors were made using the makewiggle.py script from rds files with 20 bp coverage. Then, Wig files are quantile normalized between different cell types for each histone mark or transcription factor supposing that the antibody efficiency is the same for different cell types.

All GEO datasets except some data with SoliD reads were similarly processed. For SoliD data (Ikaros, Runx1 and PolII), peak files and wig files were downloaded then, binned into 20 bp and quantile normalized.





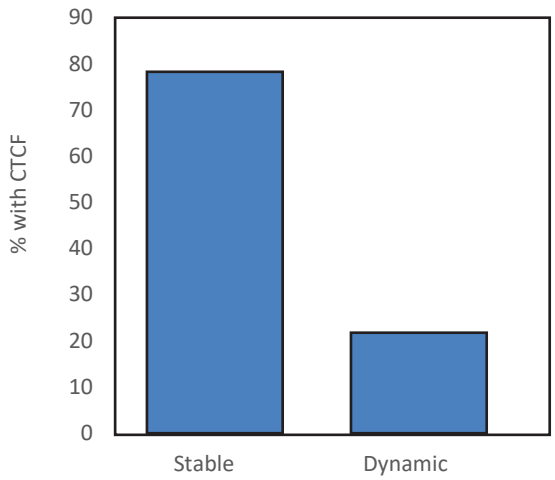


Fig 1. Conservation of TAD structure across thymocyte development.

a) CHi-C interaction heat maps for the captured region around the Il17rb gene in DN3 (left) and DP (right) cells. ChIP-seq profiles for H3K27ac and RNA polymerase II are shown below. b) Venn diagram showing exact intersections of called TAD borders across DN3, DP and ES cells. c) Bar chart showing percentages of CTCF sites found at conserved or dynamic TAD borders.

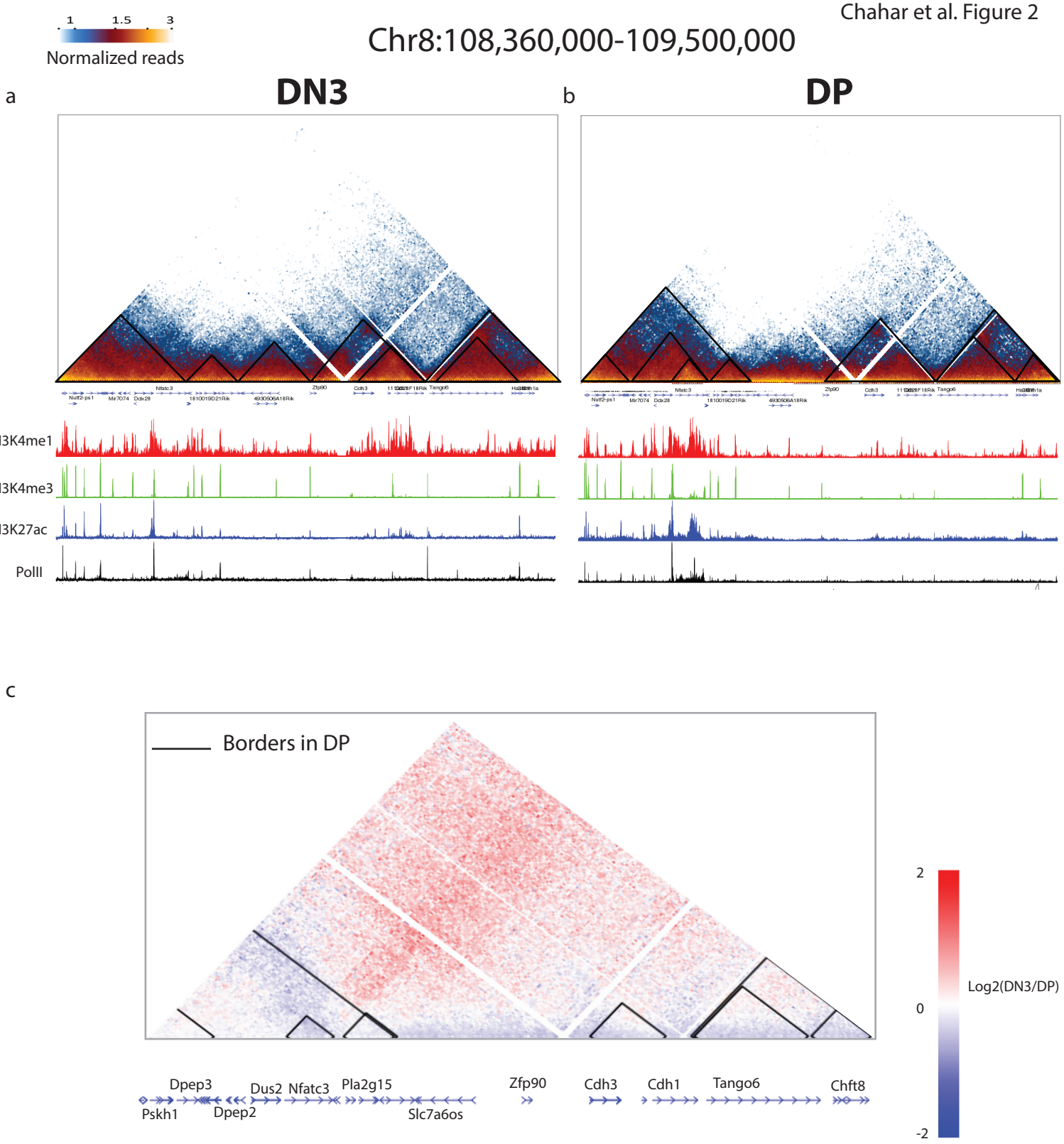
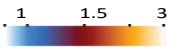


Fig 2. Transcription-coupled sub-TAD formation at the Nfatc3 gene.

CHi-C interaction heat maps for the captured region around the Nfatc3 gene in a) DN3 and b) DP cells. ChIP-seq profiles for selected epigenetic marks are shown below.
c) Heat map showing ratio of normalized DN3 CHi-C signal to DP signal for this genomic region. Nfatc3 gene forms a uniform domain of DP-enriched interactions, rather than a punctate difference in contacts at the DP-specific border.



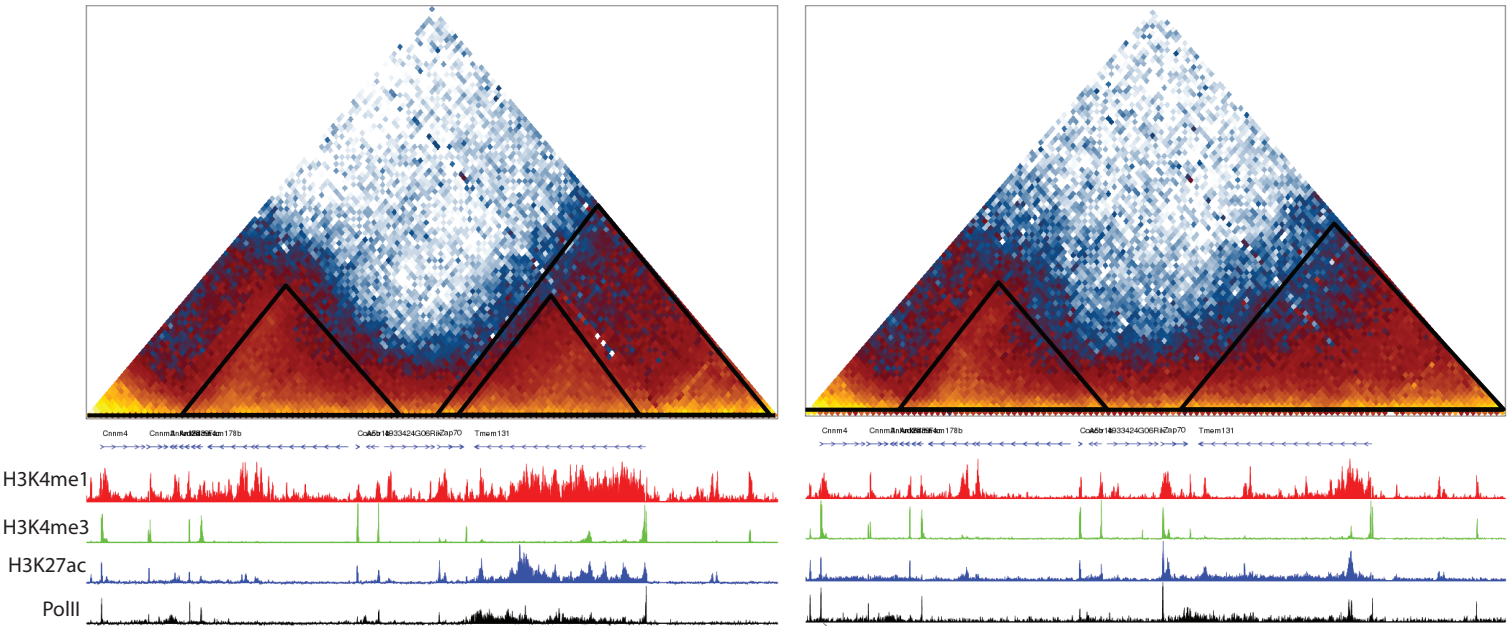
Chr1: 36515000-37110000

a

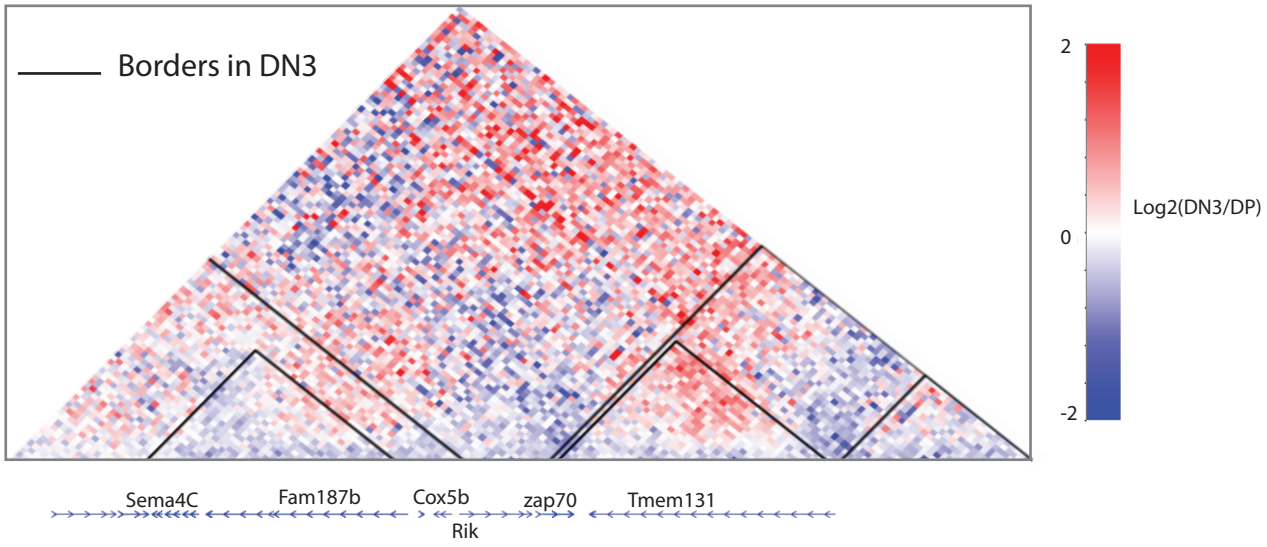
DN3

b

DP



c



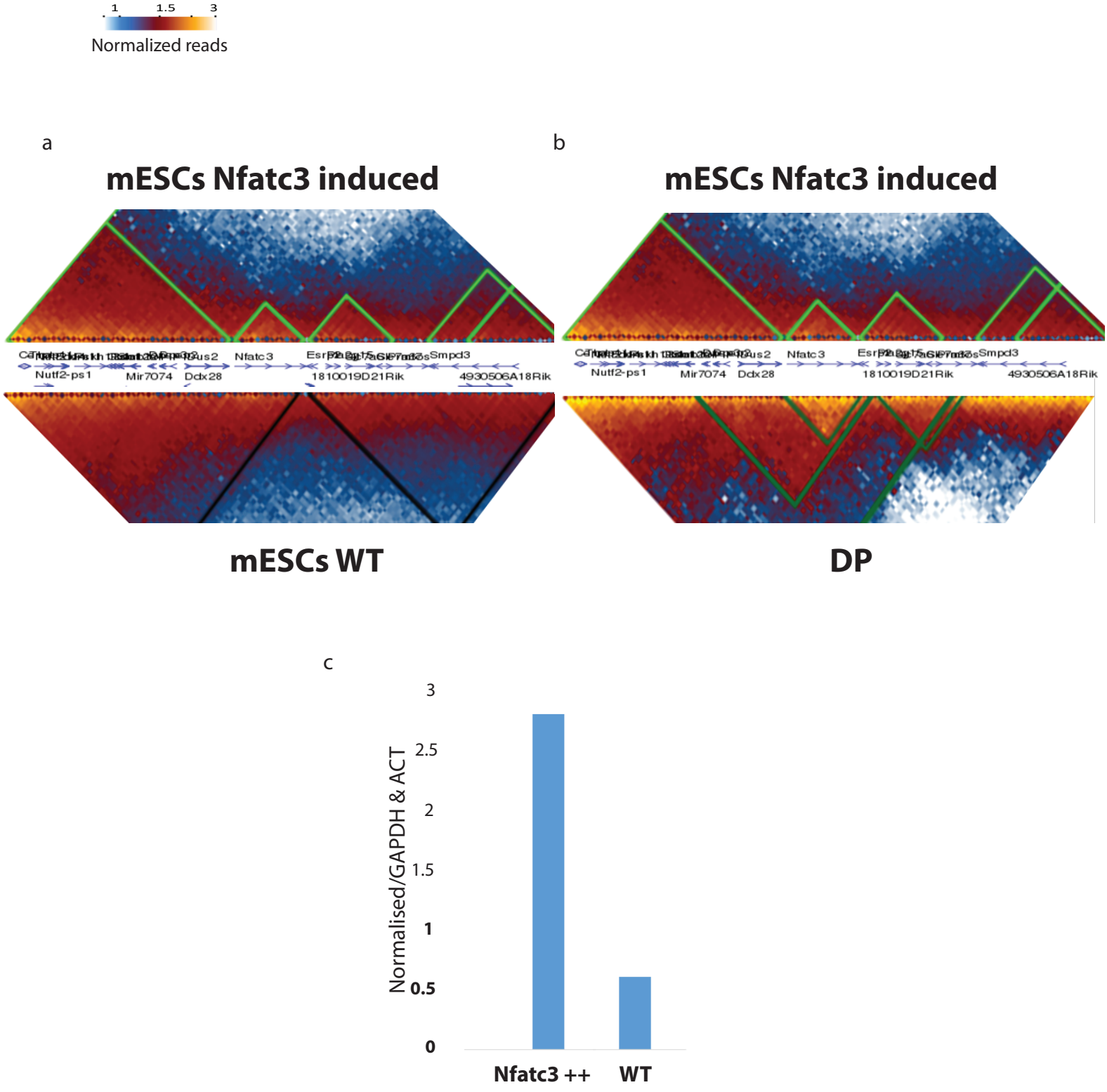


Fig 4. Transcription directly remodels the Nfatc3 sub-TAD.

CHi-C interaction maps around the same region as Fig 2, comparing a) wild-type ES cells and those with ectopic induction of Nfatc3, and b) DP cells with ES cells after ectopic induction of Nfatc3. c) qRT-PCR results for Nfatc3 expression in ES cells before and after ectopic induction.

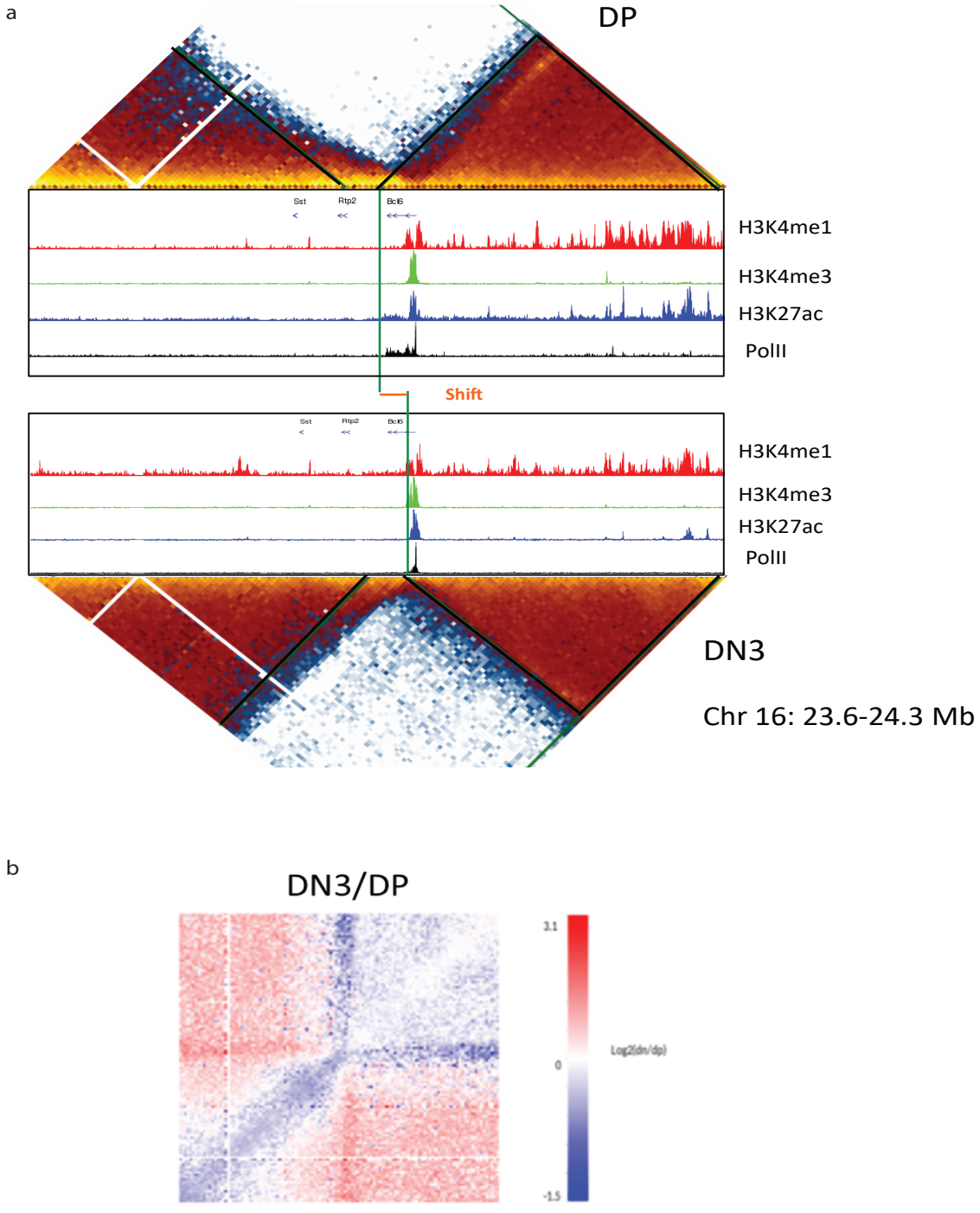


Fig 5. TAD border remodeling around the Bcl6 gene during thymocyte maturation.

a) CHi-C interactions maps for the captured region around the Bcl6 gene in DP (top) and DN3 (bottom) cells, showing an apparent border shift at the gene. Selected DN3 and DP ChIP-seq profiles are also shown. b) Heat map showing ratio of normalized DN3 CHi-C signal to DP signal for this genomic region. The border shift is clearly shown as a "stripe" of DP-increased interactions.

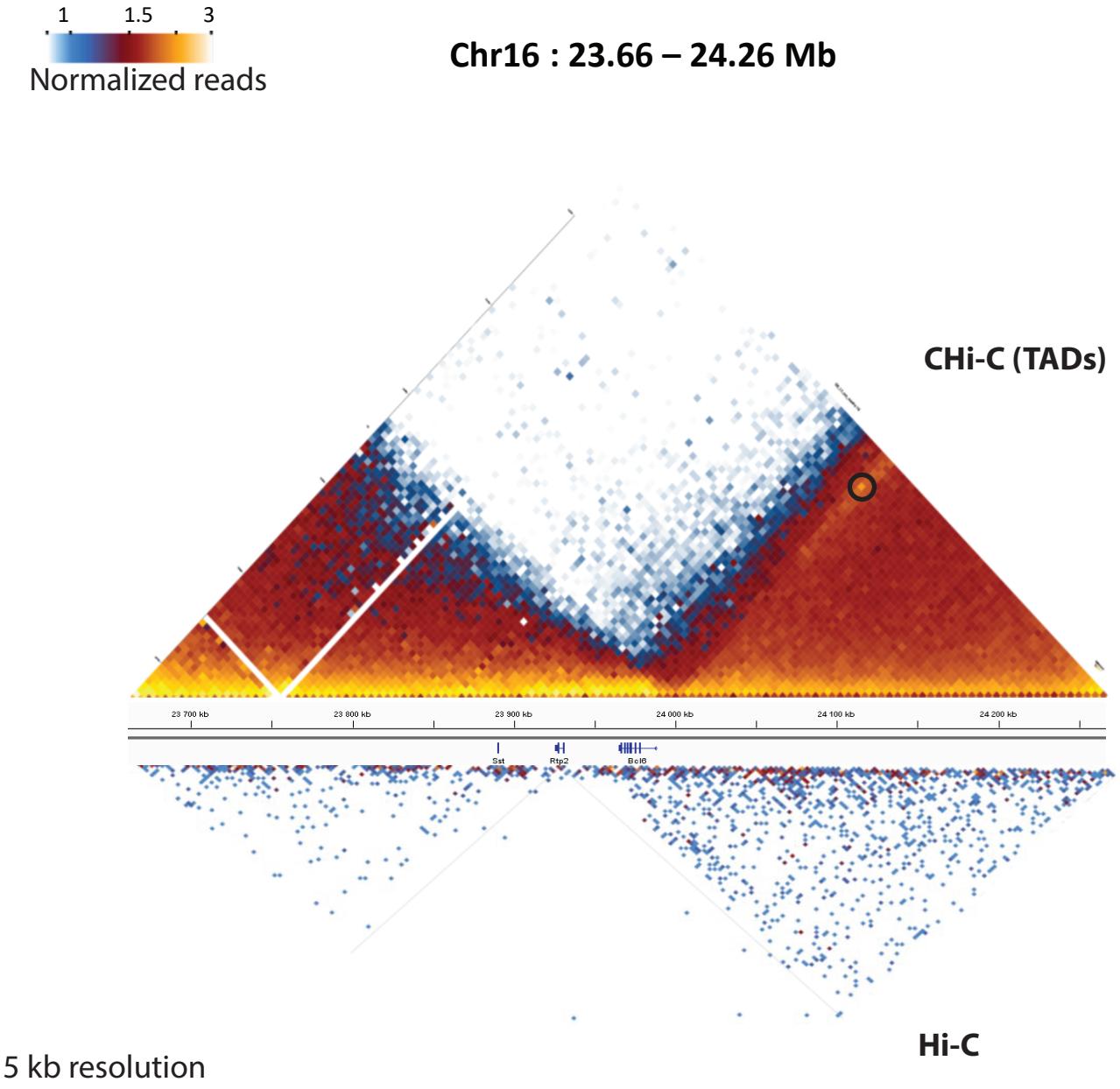
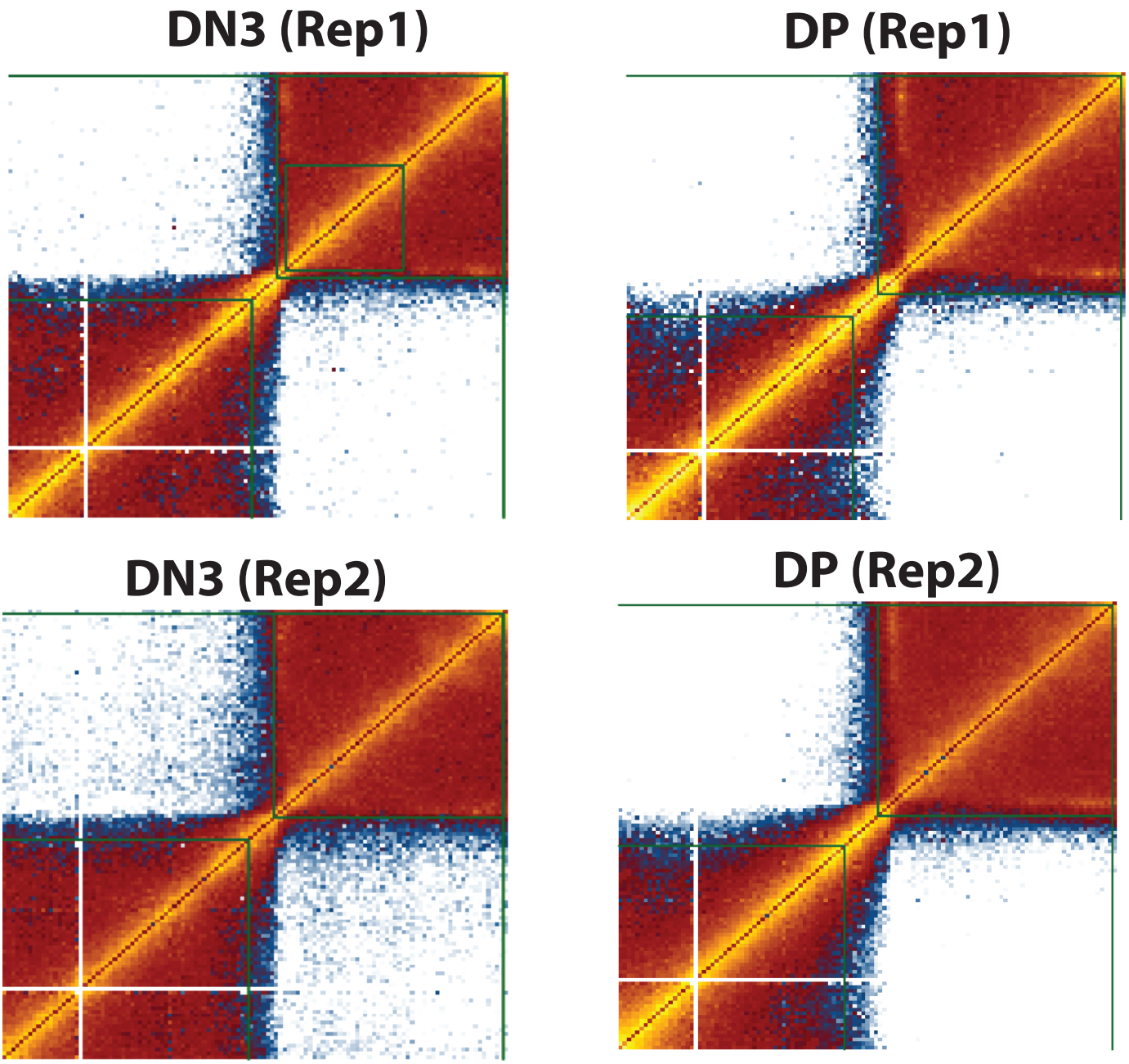


Fig S1. CHi-C enhances resolution of interaction maps.

Interaction heat maps around the Bcl6 region, plotted at 5 kb resolution from the same number of reads of a conventional Hi-C experiment (bottom) and a TAD CHi-C (top) experiment. A readily resolved enhancer-promoter interaction is highlighted by a black circle in the CHi-C experiment.



Bcl6: Chr16 (23660000 24250000)

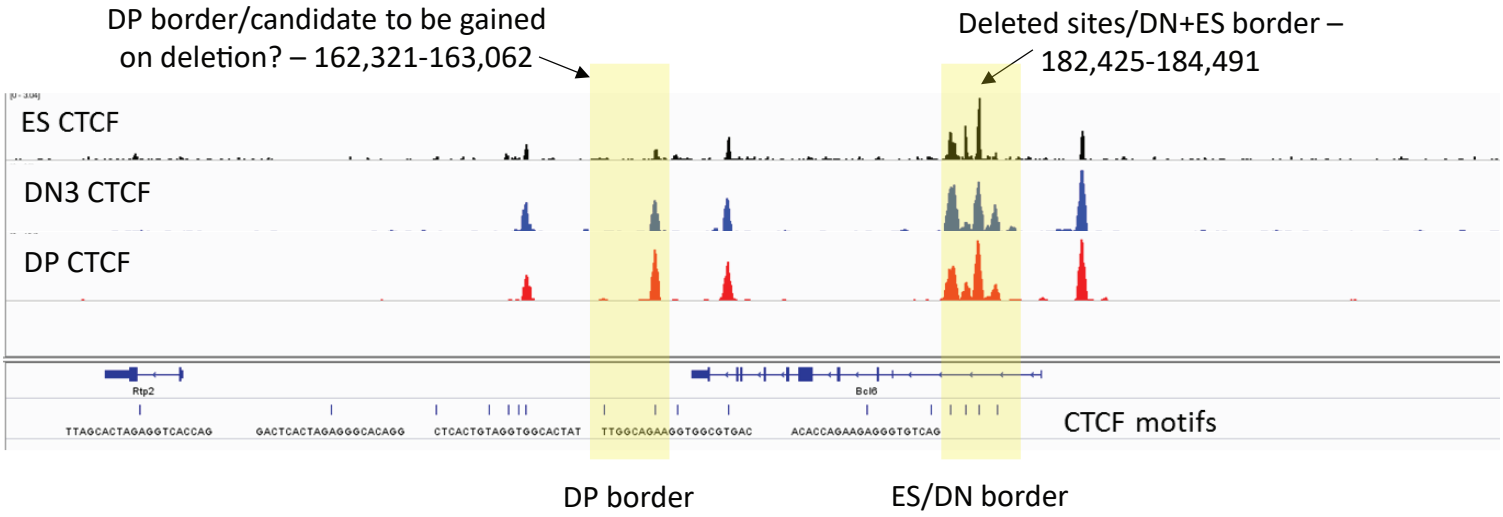


Fig S3. Potential differential CTCF binding at the Bcl6 locus.

IGV tracks for CTCF ChIP-seq at ES, DN and DP cells, around the Bcl6 region.
The DP- specific border may have quantitatively greater CTCF binding, which we are investigating by ChIP-qPCR. The position of the CTCF deletion we have performed in ES cells is also denoted, for which we will interrogate whether the TAD architecture is modified, and/or if CTCF binding is gained at the DP-upregulated site.

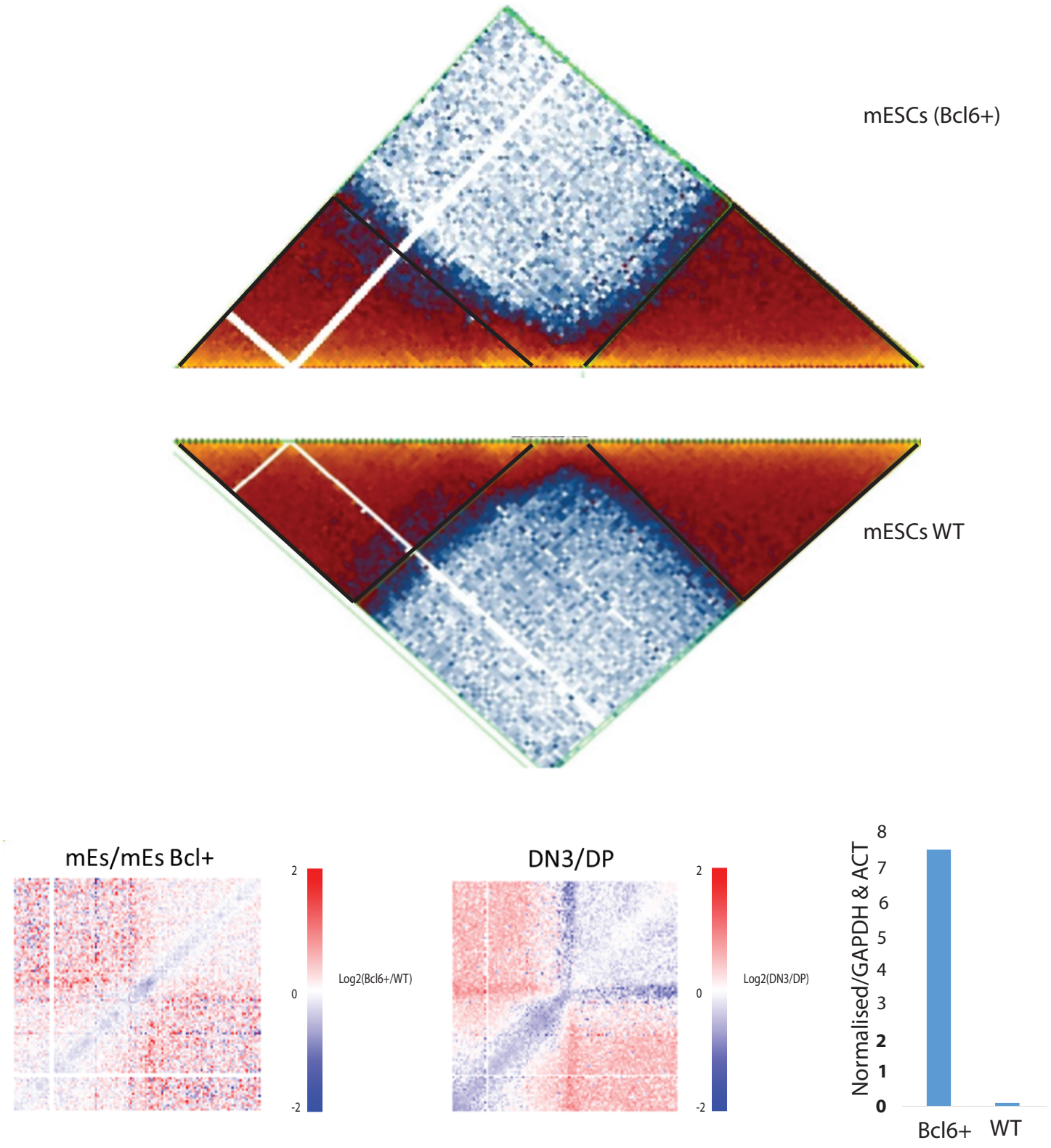


Fig S4. Transcriptional induction does not remodel the Bcl6 TAD in ES cells.

CHi-C interaction heat maps around the Bcl6 gene in ES cells before and after ectopic induction of Bcl6, showing no differences. The normalized ratios of DN3 to DP signal are also shown to highlight the lack of effect, despite qRT-PCR results showing a very strong Bcl6 induction.

Figure legends

Fig 1. Conservation of TAD structure across thymocyte development.

a) CHi-C interaction heat maps for the captured region around the *Il17rb* gene in DN3 (left) and DP (right) cells. ChIP-seq profiles for H3K27ac and RNA polymerase II are shown below. **b**) Venn diagram showing exact intersections of called TAD borders across DN3, DP and ES cells. **c)** Bar chart showing percentages of CTCF sites found at conserved or dynamic TAD borders.

Fig 2. Transcription-coupled sub-TAD formation at the *Nfatc3* gene.

CHi-C interaction heat maps for the captured region around the *Nfatc3* gene in **a**) DN3 and **b**) DP cells. ChIP-seq profiles for selected epigenetic marks are shown below. **c**) Heat map showing ratio of normalized DN3 CHi-C signal to DP signal for this genomic region. *Nfatc3* gene forms a uniform domain of DP-enriched interactions, rather than a punctate difference in contacts at the DP-specific border.

Fig 3. Transcription-coupled sub-TAD formation at the *Tmem131* gene.

CHi-C interaction heat maps for the captured region around the *Tmem131* gene in **a**) DN3 and **b)** DP cells. ChIP-seq profiles for selected epigenetic marks are shown below. **c**) Heat map showing ratio of normalized DN3 CHi-C signal to DP signal for this genomic region. *Tmem131* gene forms a more uniform domain of DN3-enriched interactions, rather than a punctate difference in contacts at the DN3-specific border.

Fig 4. Transcription directly remodels the *Nfatc3* sub-TAD.

CHi-C interaction maps around the same region as **Fig 2**, comparing **a**) wild-type ES cells and those with ectopic induction of *Nfatc3*, and **b**) DP cells with ES cells after ectopic induction of *Nfatc3*. **c)** qRT-PCR results for *Nfatc3* expression in ES cells before and after ectopic induction.

Fig 5. TAD border remodeling around the *Bcl6* gene during thymocyte maturation.

a) CHi-C interactions maps for the captured region around the *Bcl6* gene in DP (top) and DN3 (bottom) cells, showing an apparent border shift at the gene. Selected DN3 and DP ChIP-seq profiles are also shown. **b**) Heat map showing ratio of normalized DN3 CHi-C signal to DP signal for this genomic region. The border shift is clearly shown as a "stripe" of DP-increased interactions.

Tables

Gene	Known function	Expression change*	Distance to TAD border	Probes	Border position
Nfatc3	Transcription factor required for positive selection[1]	3.3 up	3 kb	3598	chr8: 108,657,000
Bcl6	Transcription factor involved in Tfh negative feedback in DP cells[2]	9.2 up	5.5 kb	1725	chr16: 23,959,000
Rag1	Recombinase enzyme at TCR loci	2.4 up	8 kb	1639	chr2: 101,497,000
Il17rb	Interleukin receptor	4.6 down	16 kb	1726	chr14: 30,793,000
Pla2g4a	Phospholipase active in thymocytes but not mature T cells	4.6 down	4 kb	1578	chr1: 151,672,000
Cdh1	E-cadherin; predominantly in thymus stroma	6 down	9 kb	3598	chr8: 109,203,000
Cd3g	Co-receptor for TCR	NS	3 kb	1925	chr9: 44,774,000

Zap70	TCR signalling protein kinase	NS	5 kb	1666	chr1: 36,813,000

Table 1. Designed regions for capture-Hi-C experiment. Capture probes are designed for 600 kb regions centred on specific TAD borders. * Mean fold expression change on transition between DN and DP cells, taken from two microarray-based experiments (Egawa and Littman, 2011; Pekowska et al., 2011). ǂ A larger (1.15 Mb) region has been designed to include both *Nfatc3* and *Cdh1*. [1] (Cante-Barrett et al., 2007) [2] (Mathew et al., 2014)

Supplementary Data

Fig S1. CHi-C enhances resolution of interaction maps.

Interaction heat maps around the *Bcl6* region, plotted at 5 kb resolution from the same number of reads of a conventional Hi-C experiment (bottom) and a TAD CHi-C (top) experiment. A readily resolved enhancer-promoter interaction is highlighted by a black circle in the CHi-C experiment.

Fig S2. High reproducibility across biological replicates.

Interaction heat maps for both biological replicates of DN3 and DP CHi-C experiments, showing that the observed TAD remodeling event at *Bcl6* is reproducible.

Fig S3. Potential differential CTCF binding at the *Bcl6* locus.

IGV tracks for CTCF ChIP-seq at ES, DN and DP cells, around the *Bcl6* region. The DP-specific border may have quantitatively greater CTCF binding, which we are investigating by ChIP-qPCR. The position of the CTCF deletion we have performed in ES cells is also denoted, for which we will interrogate whether the TAD architecture is modified, and/or if CTCF binding is gained at the DP-upregulated site.

Fig S4. Transcriptional induction does not remodel the *Bcl6* TAD in ES cells.

CHi-C interaction heat maps around the *Bcl6* gene in ES cells before and after ectopic induction of *Bcl6*, showing no differences. The normalized ratios of DN3 to DP signal are also shown to highlight the lack of effect, despite qRT-PCR results showing a very strong *Bcl6* induction.

Fig S5. TAD remodeling events uncovered in Hi-C.

Hi-C interaction heat maps for the pooled Hi-C results, showing the apparent TAD remodeling events at the *Nfatc3* and *Bcl6* regions that we had observed in CHi-C.

Discussion

General discussion and perspectives

The objectives of the thesis were to address two ambitious questions. Here, I will outline what progress was made, and the next steps I would like to take to elaborate further in the field.

- **How are chromatin configurations altered during transcriptional changes accompanying development?**

With a combination of two CHi-C strategies, one interrogating all promoter-centered interactions and the other focusing on a subset of potentially very important TADs, I was able to get a high-resolution, multi-scale view of how chromosome folding is altered during thymocyte maturation, specifically at the DN3-to-DP transition. For assessment of TADs, previous calling methods (particularly the Arrowhead algorithm for the high-coverage matrices in TAD-capture experiments; Rao et al., 2014) were suitable. However, for calling specific looping interactions from the promoter CHi-C, I was not satisfied with the available tools, and developed PromoMaxima, which I found to be more stringent and robust. As had been debated previously (e.g. Sexton and Cavalli, 2015), I found that specific chromatin looping events could be highly dynamic during development, with many interactions varying both quantitatively and qualitatively between DN3 and DP cells. In contrast, topological domains appeared much more robust to transcriptional and epigenetic changes of their component genes, consistent with a more "hard-wired" genomic architecture. However, a simplistic model of chromatin looping being completely rewired within invariant TADs also does not hold. A core of chromatin loops are maintained not just in thymocytes but also in unrelated ES cells, and we showed that a small subset of TADs may be directly remodeled by transcriptional induction. The challenge now is to identify what mechanisms, whether transcriptional or otherwise, can influence whether or not a particular chromatin architecture

can be altered, and whether this is a response to, or a driver of, functional changes such as transcriptional activity.

.1 A mixture of stable and dynamic loops during development

Like very recent promoter CHi-C studies (Hughes et al.. 2014; Schoenfelder et al., 2015a; Sahlen et al., 2015; Siersbaek et al., 2017; Freire-Pritchett et al., 2017; Rubin et al., 2017), we similarly concluded that thymocyte differentiation involves the interplay of both permissive and instructive promoter-enhancer contact networks. However, even this large network of chromatin interactions is just one small part of an even greater complexity, since many stable and dynamic chromatin loops are correlated with transcriptional downregulation of the target gene, or even with no transcriptional change at all between the tested cell types. Comparison of our CHi-C data with publicly available datasets suggest that this interactome complexity is not restricted to thymocyte lineages, and may be a general feature of all metazoan genomes. It will be particularly interesting to see if smaller numbers of cells can generate CHi-C datasets of equal quality, so that we can then apply it to characterize rarer cell types, and potentially even characterize interactomes of cancer biopsies. For example, a recent low-resolution Hi-C study claimed that the greatest nuclear architectural changes accompany the transition between DN2 and DN3 thymocytes (Hu et al., 2018). Promoter CHi-C dynamics could be very interesting to characterize what is considered the terminal step in choosing T cell fate, but DN2 populations in wild-type thymus are relatively much rarer.

A remaining technical hurdle of the CHi-C technique is that, despite a successful reduction of sequence complexity by the capture step, the method remains sub-saturating. As I have formally discussed previously, this results in poor reproducibility of interaction calling at single restriction fragments, thus limiting the resolution of the called interacting region. Previous promoter CHi-C studies have focused on the “lowest hanging fruit” of CTCF sites

and enhancers, since interacting regions can easily be functionally narrowed down to the appropriate binding site and/or peak of characteristic epigenetic mark. For other potential classes of regulatory interaction, such as distal silencers, such defining features are much less well characterized, limiting the analyses that can be performed. A wider use of promoter CHi-C strategies with more frequently-cutting restriction enzymes (e.g. Joshi et al., 2015; Sahlen et al., 2015) may improve on the resolution, but it remains to be seen systematically whether the increased complexity of the pool of possible ligation products actually creates more of a problem than it solves. Other options to improve on the resolution are to perform more and more systematic 4C experiments, and/or to sequence (C)Hi-C libraries deeper and deeper. Another means to functionally home in on the potential regulatory potential of interacting regions is to perform high-throughput reporter assays, such as STARR-seq for enhancer screening (Arnold et al., 2013) and adapt them to read out other functional elements. For example, a STARR-seq-like approach has been used to assess promoter responsiveness (Arnold et al., 2017); a successful use of an analogous approach to identify silencers at higher resolution could be invaluable in identifying the epigenetic features controlling their looping to target promoters and/or mediating transcriptional repression.

.1.1 Enhancer-promoter communication: when to loop?

Despite their relative ease of epigenetic analysis compared to other regulatory factors, there is still no apparent “rule” dictating when an enhancer contacts its target promoter. Although the exact numbers can vary depending on the arbitrary thresholds that are set, I identified a seemingly equivalent number of instructive and permissive enhancer-promoter contacts. Despite a large number of studies dissecting the epigenetic hallmarks of “poised” versus “active” enhancers, both types appear equally as likely to contact gene promoters. The one study to date that has claimed to distinguish features at permissive and instructive loops

identified cohesin as being enriched at stable interactions, and uncovered specific transcription factors at induced loops (Rubin et al., 2017). No cohesin profiles currently exist for DN3 cells, so I was unable to directly compare this finding. However, hierarchical clustering analysis did not reveal any transcription factor combinations that could predict loop timing. I would like to attempt the ROC approach that Rubin et al. used to obtain their findings, but it is more likely that many different factors play a context-dependent role in modulating the loop at small subsets of genes, confounding global analysis.

.1.2 A LINE to transcriptional silencing?

Within the previously described technical limitations of CHi-C analysis, I was able to identify a number of putative distal silencer elements, which were subsequently functionally validated by other members of the group. This could represent a previously underappreciated class of gene regulatory element, but the results of ongoing CRISPR deletion experiments are required to assess whether these interactions are functionally meaningful. I anticipate that such silencer interactions, if functional, are likely to play roles in fine-tuning developmental gene expression, rather than be absolutely required to prevent aberrant expression of potentially dangerous genes, since the latter case is not evolutionarily robust. In fact, our growing appreciation of most enhancers is to similarly provide robust gene control in concert with other regulatory elements (Osterwalder et al., 2018). Despite this candidate list of silencers not having many clear epigenetic characteristics, apart from a depletion of active histone marks, I was able to extract an interesting feature, which has been partially validated in luciferase assays: an enrichment for LINEs juxtaposed to a CTCF motif facing the target promoter. Ongoing deletion experiments will try to assess and distinguish the roles of the LINE and the CTCF motifs in chromatin looping and/or gene silencing, but our favored hypothesis is that repressors brought to the LINE as part of the host's genome defenses are

co-opted to inhibit target gene transcription. The publicly available H3K9me3 ChIP profile exists for whole thymus rather than specific thymocyte subsets, and in any case is very difficult to analyze robustly due to the mark existing as broad, weak domains, and its prevalence on repetitive regions that are poorly mapped. I anticipate that if we can analyze more cell type-specific H3K9me3 profiles in a more robust manner, we may see a greater prevalence at *bona fide* distal silencers, especially on TEs.

Even if a functional proof of principle is obtained for LINEs as distal silencers, the next challenge in understanding their potential role in development is to see if and how their action is modulated during developmental transitions. It may suffice for positive transcriptional signals to overcome these potentially weak repressive interactions; we have preliminary evidence suggesting that silencer interactions are lost on ectopic induction of the gene. Alternatively, any truly regulated silencers may contain additional sequences conferring developmental control, or the binding of CTCF (and presumably, ability to confer a chromatin loop interaction) may be developmentally regulated at these elements. Future perturbation experiments will be required to explore this interesting avenue of research.

.1.3 Looping beyond transcriptional control?

Curiously, the majority of dynamic promoter interactions at the DN3-to-DP transition (and also when comparing thymocytes to ES cells) appeared unrelated to transcriptional control. Many of these dynamic interactions are as robust and cell type-specific as “functional” enhancer interactions, so are unlikely to result as a simple technical problem of the CHi-C method. A few possible explanations may account for some of these dynamic interactions, but the majority appear a genuine mystery:

- Poised enhancers, which will actually play a functional role at a later developmental stage (or played one at an earlier stage). Similarly, the gene may be swapping usage of

very cell type-specific enhancers, with no net result in expression level. Clustering within these classes of interactions identified some enhancer hallmarks in a minority of cases.

- Developmentally plastic CTCF-mediated loops. However, the functional significance of these is unclear.
- Genes at a TAD border swap between the two adjacent TADs, as has been described for Hox genes during development (Andrey et al., 2013).

1.2 TADs: an architectural buffer?

Unlike chromatin loops, we have shown that the majority of TADs are robust to transcriptional changes during development. Although we confirmed that most TADs are conserved (Dixon et al., 2015), we observed, like other recent studies (e.g. Bonev et al., 2017), specific genomic contexts where transcriptional induction or upregulation is correlated with spatial chromosomal remodeling. Our studies and others have identified three different classes of TAD remodeling event: creation of a border at the TSS of an activated gene (Bonev et al., 2017); generation of a sub-domain comprising the activated gene body; shifting of a TAD border to accommodate the fully transcribed gene. Although few ectopic induction experiments have been performed, to date only the second type of remodeling event has been shown to be caused directly by transcription. More systematic studies are required to see if this is indeed the case, to what extent transcription can really remodel domains, and whether blocks to remodeling impede transcriptional activation.

In any case, the large number of counter-examples of differentially expressed genes which do not display TAD architectural differences, even when studied at high resolution, suggest that most TADs are indeed invariant to gene expression differences. Two major questions remaining are what causes the difference between plastic and rigid TADs, and are

rigid TADs necessary for gene control? For the latter, TADs have been proposed to comprise an architectural buffer, either limiting the functional range of potentially dangerous enhancers (Lupianez et al., 2015), and/or conversely limiting the search space for efficient action of enhancers on their cognate genes (Symmons et al., 2016). Further perturbation studies are likely to solidify this hypothesis.

Hi-C studies are converging on identifying TADs as largely stable structures, but these are ultimately experiments capturing population average snapshots of fixed nuclei. If TADs are truly buffers of genomic activities, we would expect them to be stable structures in all nuclei, all throughout interphase, something that has yet to be demonstrated. Improvements in live imaging experiments put us in an exciting place to address this question. In the context of my own findings, I would be very interested to see if the remodeling events of the developmentally more malleable TADs can be visualized, and to see whether they precede or occur concomitantly with transcriptional activation.

- **Is chromatin topology important in controlling cell differentiation and development?**

Hi-C experiments alone can only identify chromatin interactions; follow-up studies are necessary to demonstrate any functional relevance. A major technical challenge to address the causal nature of chromatin architecture in transcriptional control is to perturb chromatin architecture specifically and in a way decoupled from indirect effects on transcription. For example, initial studies deleting transcription factors could not decouple the factor effects on chromatin looping and direct transcriptional regulation at the promoter (Vakoc et al., 2005). Seminal studies in the beta-globin locus showed that chromatin looping could be induced by protein dimerization events, and that this in itself was sufficient to activate transcription (Deng et al., 2012). Subsequently, it has been shown that completely artificial protein dimerization systems can induce chromatin loops at different loci and cell types (Morgan et

al., 2017). The universality of this approach has yet to be explored, but this tool and its variants has the potential to allow promoter-enhancer and promoter-silencer loops to be specifically switched on and off at key developmental timepoints, and their functional output read as changes in gene expression. I would be very interested to see the results of induced looping during thymocyte development, although for technical reasons, these experiments may be limited to less useful cell lines, or at least adapted to more technically feasible studies of ES cell differentiation.

Overall, the CHi-C datasets that I have analyzed have uncovered an extremely rich network of stable and developmentally dynamic chromatin architectures at multiple scales, of which at least a subset appear important for transcriptional control. These data are likely to inform myriad perturbation studies to uncover the potential role of "established" (i.e. enhancer-promoter loops) and novel regulatory interactions in controlling appropriate developmental gene expression.

Materials and Methods

This chapter is organized as the results chapter.

I. Hi-C and CHi-C quality control

Hi-C and CHi-C datasets are initially processed the same way, with downstream analyses catered to the specific questions asked in each application (**Fig 1**). CHi-C datasets then require an additional filtering step to class the interactions as non-captured, single bait-captured or double bait-captured ones.

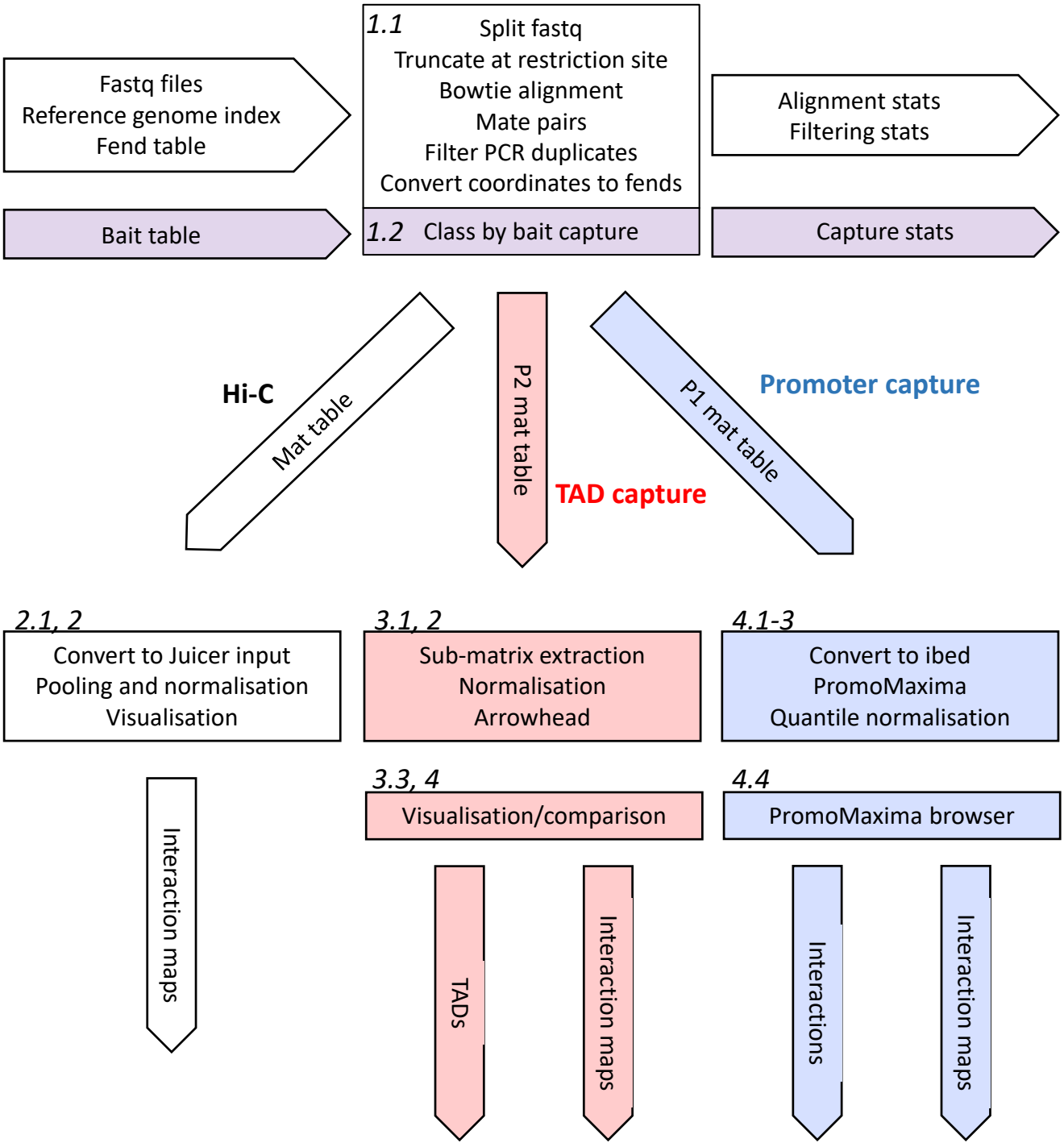
1. Universal processing and filtering

The pipeline used is very similar to that of Sexton et al., 2012, or described in HiCUP (Wingett et al., 2015). The required inputs are the fastq files from the sequencing reaction (separate files for each end of a paired end read), the Bowtie index file for the reference genome (mm9; Langmead et al., 2009), and a table describing all the restriction fragment ends (fends) within the genome. Fend tables for mm9 with HindIII or DpnII digestion were already available in the group, and are of the format: *fend* (unique integer ID), *frag* (integer ID), *chr*, *coord*, *strand* (+ or -).

Custom perl and R scripts perform the following pipeline, with tasks run in parallel on a cluster for efficiency:

- Fastq files split into multiple files of 500,000 reads for parallel alignment.
- Recognition of Hi-C ligation junction sequence and truncation, so that only sequence within a single restriction fragment is input for alignment. Truncation statistics are automatically generated.
- Alignment to the reference genome with Bowtie. Only unique alignments are kept, with the parameters -m 1 --best --strata. Mapping statistics are automatically

Fig 1


Fastq files
Reference genome index
Fend table
1.1
Split fastq
Truncate at restriction site
Bowtie alignment
Mate pairs
Filter PCR duplicates
Convert coordinates to fends
Alignment stats
Filtering stats
Bait table
1.2
Class by bait capture
Capture stats
Hi-C
Mat table
P2 mat table
TAD capture
P1 mat table
Promoter capture
2.1, 2
Convert to Juicer input
Pooling and normalisation
Visualisation
3.1, 2
Sub-matrix extraction
Normalisation
Arrowhead
4.1-3
Convert to ibed
PromoMaxima
Quantile normalisation
Interaction maps
3.3, 4
Visualisation/comparison
4.4
PromoMaxima browser
TADs
Interaction maps
Interactions
Interaction maps

generated, and the aligned reads are output in the standard Bowtie format: *read_ID, strand, chr, coord, sequence, Phred quality, specific alignment notes*.

- Pairs are mated, based on the corresponding read_IDs of the files for reads of each end. In other words, only reads where both ends of the paired-end read are successfully and uniquely aligned are kept. Output of paired file is: *chr1, coord1, strand1, chr2, coord2, strand2*.
- Removal of PCR duplicates. Any lines of the paired file where both pairs of the reads are exactly identical to that from another line are filtered out. Pairing and deduplication statistics are automatically generated.
- Conversion to fend space. The chromosome/coordinate positions of the pair files are converted to their corresponding fends, using the fend table as a look-up. At this stage, certain erroneous reads are filtered out. First, reads where the sequenced tag falls exactly on a restriction site are removed, since these correspond to non-ligated ends rather than randomly sonicated regions, which have been shown to add noise to the interaction maps (Sexton et al., 2012). Secondly, paired reads where both ends fall on the same fragment are filtered, as these correspond to self-circularisation events. Third, paired reads that are facing each other and are separated by less than a threshold distance (2 kb) are filtered, since they could possibly result from contiguous, non-digested genomic regions. Fourth, the theoretical size of the sequenced fragment is inferred from the positions of the sequenced tags and the restriction sites; if this is larger than a threshold size (500 bp), the reads are filtered since they may represent sequencing errors. The other events are merged together and output to a mat table format: *fend1* (fend unique ID), *fend2, count*. The filtering statistics, as well as the read cis/trans ratios are automatically generated.

To identify the following statistics (Alignment: % Truncated, % Mapped, % Paired and % Unique; Filtering and Cis/Trans ratio), as shown in Fig 1, datasets are pooled as following:

* 2 Hi-C libraries with two different restriction enzyme (HindIII and DpnII) are generated for both DN3 and DP cells. All of these datasets are pooled together to identify their corresponding statistics (Total number of reads: 1,531,228,478).

* Two biological replicates with a technical replicate for each are produced for CHi-C (promoters) for both DN3 and DP. In total, 8 CHi-C (promoters) libraries (Total number of reads: 4092213841) are pooled together for these statistics.

* Two biological replicates for both DN3 and DP are generated for CHi-C (TAD borders). All these 4 libraries are pooled together (Total number of reads: 1960056746).

2. Processing captured reads from CHi-C

A modification to the above pipeline has been added to account for interactions that are from non-captured (P0), single-captured (P1) or double-captured (P2) CHi-C paired-end reads, and is applied when all the split files have been converted to fend space and are merged at the very end into one mat table. A custom perl script is used, which requires a modified fend table, identical to that described above but with the extra column *Probe*, which has the value 0 for fends not covered by a capture probe, and 1 for fends that are covered by a probe. Instead of creating a single mat table as for Hi-C results, this script creates three separate mat tables, for P0 (both interacting fends are 0), P1 (one of the two interacting fends is 1), and P2 (both interacting fends are 1). Capture efficiency statistics are automatically generated.

2 Capture efficiency control

To determine the capture efficiency of both experiments (CHi-C promoters and TADs), we first identify the different populations of chimeras (P0, P1, P2; see Results I.2). In Fig2A and

B, the percentages of the different populations were computed for each of the DN3 and DP replicate libraries (see above) of CHi-C (blue) or corresponding Hi-C with the same restriction enzyme (red), and their mean and standard deviations are plotted as bar plots.

The capture efficiency of each individual probe used in the CHi-C promoter experiments is calculated as follow: the number of captured reads containing that specific probe (P1 or P2)/ Total number of valid reads in the dataset. To fairly compare the capture efficiency of each probe between different cell types, we also used published CHi-C promoters (mEScs and FL). The distribution of these probe capture efficiencies were plotted as box plots (Fig 2C). To assess any potential cell type variability in individual probe capture efficiencies, Fig 2D shows the Spearman correlations between these capture efficiencies for pairwise combinations of the CHi-C datasets. The "Random" dataset was generated by performing CHi-C steps on purified genomic DNA (i.e. assessing the capture efficiency on randomly ligated restriction fragments, since all effects of proximity on the linear chromosome fiber should be eliminated).

3. Construction of "pooled" Hi-C contact matrices and visualization

The mat tables (See Results I.1.2) were sequentially converted by a custom perl script into a format compatible with the Juicer Pre tool: *read name, strand1, chr1, coord1, fragment1, strand2, chr2, coord2, fragment2* (the lines are replicated the same number of times as there are reads for that particular pairwise combination). The concatenated file is then sorted and input into Juicer Pre (Durand et al., 2016), which converts to a single file in the binary .hic format. The .hic output of Juicer Pre is directly visualized with the Java tool Juicebox (Durand et al., 2016).

4. Hi-C/CHi-C (TADs) matrices normalization

In order to remove any biases from Hi-C/CHi-C matrices, we use matrix balancing method for matrices normalization. We use KR, which is a derivative algorithm of matrix balancing, for Hi-C matrices. This is accomplished using Juicer pre command included in Juicer pipeline. For CHi-C (TADs) we use ICE, which is also a matrix balancing derivative method. The ICE algorithm is implemented in a custom R script applied on selected sub-matrices of CHi-C (TADs).

II. PromoMaxima: a pipeline for detection and visualization of cis-DNA looping in Capture Hi-C

A full documentation of the PromoMaxima Package is found in this link: https://github.com/yousra291987/ChiCMaxima

III. Developmentally dynamic gene promoter interactions in transcriptional activation and repression

1. Identification of chromatin loops using PromoMaxima

1.1 Generating input files for PromoMaxima

From P1 mat tables (corresponding to promoter bait interactions with a non-promoter, non-captured region), a custom perl script converts the data into an ibed file of the following format: *ID_fragment1, Chr_fragment1, Start_fragment1, End_fragment1, Gene name, ID_fragment2, Chr_fragment2, Start_fragment2, End_fragment2, Number of reads*.

1.2 PromoMaxima

The PromoMaxima R scripts is described in the accompanying manuscript, which takes the ibed input and outputs a file of the same ibed format for the subset of interactions that are called as hits. Throughout this thesis, all PromoMaxima analyses were performed with the following settings: -w =50; -s=0.05, -d=30000.

1.3 Quantile normalization for comparison of biological samples

For each bait separately, the interaction scores from different datasets (for fragments within 1.5 Mb of the bait) are quantile normalised using the normalizeBetweenArrays function from the R package limma. New ibed files are generated with an extra column for the quantile normalised scores for each dataset.

1.4 Visualization on PromoMaxima browser

We use the browser implemented in PromoMaxima pipeline to visualize the CHi-C (promoters) data. All figures produced are screenshots produced from this browser.

2. Epigenomic analysis

2.1 Dataset sources

All dataset were downloaded from GEO database (Results III: Table S2).

2.2 ChIP-seq data processing

I aligned Chip-seq raw reads (from GEO) to the mm9 genome using bowtie2. Then, I called peaks using ERANGE V4.0 (http://woldlab.caltech.edu/rnaseq/). Mapped reads in the SAM file are first transformed into native Erange reads store .rds file. Then, peaks are identified with the peaks finder tools in Erange with –nodirectionality and –notrim parameters. Erange returns a per-peaks p-value. Wig files of Histone marks and transcription factors were made using the makewiggle.py script from rds files with 20 bp coverage. Afterwards, Wig files are quantile normalized between different cell types for each histone mark or transcription factor, using the normalizeBetweenArrays function from limma. The quantile normalised wig files are the inputs in all browser shots shown in this thesis, which are autoscaled in the browser.

For SoliD datasets, which did not have raw reads available (Ikaros, Runx1 and PolII), peak files and wig files were downloaded directly from GEO, binned into 20 bp bins and quantile normalized with limma.

2.3 RNA-seq data processing

RNA Seq data for DN3, DP and mES from GEO database (Results III: Table 2) were downloaded as fastq files. Reads were mapped to the mm9 genome using bowtie2. Mapped

reads in SAM format are then transformed into rds file by using ERANGE V4.0 tools (makerdsfrombowtie.py). For each gene, ERANGE counts unique reads falling on the gene models using rpkm normalization. The output is a text file with each line corresponding to a specific gene with its corresponding rpkm value. These genes are then classed into five groups: 0 = 0 rpkm; 1-4 = first to fourth quantiles of rpkm values of the remaining gene set.

2.4 Enrichment analysis

The enrichment for chromatin marks and transcription factor in interacting fragments was calculated using the proportion of fragments that overlap with a peak for the chromatin mark or transcription factor, divided by the proportion of all non-bait fragments that overlap with such a peak. Then, resulting values were converted to its log2 value, so that positive values represent an enrichment compared with all non-bait fragments and a negative value represents depletion.

To assign interacting fragments to an expression class, the interacting fragment must interact only with baits from the same expression class, otherwise it is excluded from the analysis.

3. 4C-seq analysis

Custom perl and R scripts are used in the 4C-seq analysis pipeline (de Wit et al., 2015), which comprises:

- Fastq files are demultiplexed into the reads from specific baits by the sabre tool (https://github.com/najoshi/sabre), using the 4C primer sequence and expected sequence up to the DpnII restriction site as the (long) barcode.
- Demultiplexed reads not starting with the expected DpnII site (GATC) are filtered out.
- The reads are mapped to the mm9 genome with bowtie, then converted to fend space, essentially as for the Hi-C pipeline. This latter step automatically filters out reads that comprise more than 2% of the total reads, which are predominantly non-digested bait-linked sequences, and (rarely) exceptional PCR duplication artefacts. All

intrachromosomal interactions are automatically output as a bedgraph file, including all non-covered restriction fragments.

- These bedgraph files are simultaneously smoothed by a running mean (from the zoo R package) and quantile normalized by limma, and the output bedgraph files are directly plotted in the IGV or PromoMaxima browsers.

4. Classes definition (A-F)

Using arbitrary cutoff (1 for log2 strength of interaction and 1 for log2 gene expression), we classify the significant interactions into different cis-regulatory elements according to their corresponding gene expression.

For each cell type (DN3/DP), we apply a cutoff (1) on both gene expression and strength of interaction only in one cell type, according to that we call:

*A = both strength of interaction (log2 fold change) and the gene expression (log2 fold change) > 1

*B = strength of interaction < 1 and the gene expression > 1

*C= strength of interaction >1 and the gene expression <-1

*D= strength of interaction < 1 and the gene expression < -1

*E = strength of interaction < 1 and -0.25 <the gene expression<0.25 (~0)

*F= strength of interaction >1 and -0.25 <the gene expression<0.25 (~0)

IV. TADs caller benchmarking

We used different tools for calling TADs in the TADs CHi-C experiment:

*TADbit (Serra, Baù, Filion, & Marti-Renom, 2016) with these parameters: TAD size= the entire chromosome (default); identify_centromeric_regions=TRUE

*The insulation score (Crane et al,2015) : the insulation square=50kb; the insulation delta span=20kb; noise threshold=16; boundary margin of error=0.1

*Armatus (Filippova, Patro, Duggal, & Kingsford, 2014): gamma_max=0.05; All other parameters=default

*Arrowhead (Durand et al, 2016): K=None; R=5Kb; All others=default

The matrices used in this chapter are submatrices of double captured regions (resolution 5Kb) extracted using a custom perl script then normalized using R script for matrix balancing (Grubert et al,.. 2015) (See above: I.4).

V. Transcription directly remodels a small subset of topologically associated domains

1. Sub-matrix extraction and normalization

A custom perl script converts the P2 mat table into an equivalent table with fixed bins of 5 kb width. The relevant sub-matrices are then extracted by a simple filter for chromosome and coordinate range, then input into a custom R script for matrix balancing (Grubert et al,., 2015).

2. Arrowhead TAD calling

The normalized sub-matrices from *3.1* are converted to a .hic format as for *2.1*, then input into Juicer (Durand et al., 2016). From this, the Arrowhead algorithm (also within the Juicebox suite) is applied with the parameter -r 5 kb.

3. Visualization of sub-matrices

In R, the normalized sub-matrices (5 kb bins) are plotted as heat maps, and the output rectangle format of Arrowhead is overlaid to plot the called TADs. All TAD capture heat matrices plotted in the thesis use the same colour scheme: 2=”white”, 5=”dodgerblue4”, 10= “darkred”, 3 0=”orange”, 50 =”yellow ”, 80=" lightgoldenrodyellow".
The break values represents the quantiles (0-25-50-75-90-100) of normalized values multiplied by 1000.

4. Comparing sub-matrices

The normalized interaction values of one matrix are divided by another, and the log2 of this ratio is computed. These values are plotted as heat maps, just like *3.3*, with the different color scheme, used in all comparative plots in this thesis: -2= “blue”, 0=”white”,2=”red”.

Figures Annexes

Introduction

Fig 4. All Hi-C datasets from DN3 thymocytes were pooled (see *2.1*) and then visualized with Juicebox (see *2.2*) at different scales. Selected architectural features were annotated manually.

Hi-C and CHi-C Quality Control

Fig 1. Statistics are directly recovered from the Hi-C processing pipeline (see *1.1*). The mean percentages for all Hi-C and CHi-C datasets generated in the group are plotted with their standard errors *Hi-C and CHi-C Quality Control,* ***Fig 2***. A and B) Percentages are directly output from the capture processing pipeline (see *1.2*); mean percentages are plotted with their standard errors. C) The capture efficiency of each individual probe is calculated as (number of CHi-C reads containing the probe / total number of CHi-C reads) and the distributions for each cell type are plotted as a box plot. D) The Pearson correlation coefficients for individual probe capture efficiencies are computed for each pairwise combination of CHi-C datasets. The "Random" dataset is derived from performing the entire CHi-C protocol on genomic DNA, which has been digested and re-ligated under non-dilute conditions to generate a random mix of 3C ligation products, which should have no dependence on genomic distance.

PromoMaxima: a pipeline for detection and visualization of cis-DNA looping in Capture Hi-C

Fig 1. PromoMaxima is applied to mES CHi-C data. Specifically, all interactions with the *Nxt1* bait and that are within 1.5 Mb of the bait are extracted. A scatter plot is made of raw

CHi-C read counts against genomic distance from bait. The smoothed line is a plot of the loess fit (span = 0.05), using the same y-axis scaling. The dotted line is a plot of the negative binomial-fit to the distance decay.

Fig 2. PromoMaxima is applied to mES CHi-C data for the *Hoxa5* bait, as previously. The loess fit curve is shown with the PromoMaxima-called interactions (dotted line) and alongside the corresponding mES *Hoxa5* 4C-seq profile (see *6*).

Fig 3. A) Interactions are called for the same mES CHi-C dataset by GOTHiC, CHiCAGO and PromoMaxima, and compared in Venn diagrams. B) The GOTHiC (left) and CHiCAGO-called (right) mES interactions are classed as those that are called by PromoMaxima and those that are not, and the distributions of their GOTHiC (-log p-value) and CHiCAGO (weighted probability converted to a score, as in Cairns et al., 2016) interaction scores are shown in box plots. The difference between the two classes is assessed by Wilcoxon rank sum test. C) The enrichments in called interactions for CTCF or certain histone marks are assessed for the interactions called in mES CHi-C data by the three different methods (see *5.4*).

Fig 4. Features of the PromoMaxima browser will be demonstrated during the thesis defence.

Fig S1. Jaccard indices for interactions called within biological replicates are shown for mES, DN3 and DP CHi-C datasets, with distributions plotted for interactions called by different methods.

Fig S2. ROC curve produced for different sets of parameters of PromoMaxima. One thousand viewpoints were subsampled, and the interactions were called by PromoMaxima using different parameters. All non-bait restriction fragments within these analysed regions (within 1.5 Mb of the different viewpoints) were then classed as a "hit" or "non -hit" based on

their PromoMaxima call, and were compared in a ROC analysiswith the R package RORC, using the CHi-C read counts as the corresponding values.

Fig S3. For mES CHi-C data, the coordinates of the interacting regions are obtained from each replicate individually by PromoMaxima. The closest bait-matched interaction in the second replicate is found for each interaction called in the first replicate, and the distance distribution is plotted as a histogram. The line denotes the frequency density.

Developmentally dynamic gene promoter interactions in transcriptional activation and repression

The mES TAD coordinates are taken from Bonev et al., 2017; the PromoMaxima-called CHi-C interactions are classed as being intra- or inter-TAD depending on whether a TAD border is present in between the bait and interacting region. Interactions where either bait or interacting region fall exactly on a TAD border are removed from the analysis. The p-values for bias towards facing CTCF sites within interacting regions is calculated as follows. First, interactions are filtered to only contain CTCF ChIP-seq peaks (see *5.2*) falling on a recognisable CTCF motif (taken from PWMTools; Ambrosini). Assuming an equal probability of each motif facing in either orientation, the p-value for the number of motifs facing the bait to be equal to or greater than the observed value is calculated directly from a binomial distribution.

Fig 1. A) PromoMaxima-called interactions for DN3, DP and mES CHi-C data are compared in a Venn diagram. B and C) Genes are classed by RNA-seq data from DN3 (0 = 0 RPKM; 1-4 = first to last quartiles of RPKM values) (see *5.3*), and box plots are shown for distributions of B) numbers of called interactions; C) distance between bait and interacting region. D) Percentages of DN3 bait-interacting regions containing a ChIP-seq peak for particular factors or histone marks (see *5.2*). E) Enrichments of the same factor/mark peaks in DN3 (blue) and DP (red)-called interactions (see *5.4*). F-H) Quantile normalized (see *4.3*) CHi-C plots for the

region surrounding the *Ikzf1* bait is plotted alongside corresponding H3K27ac ChIP-seq profiles (see *5.2*) and 4C-seq profiles (see *6*). The called interaction, conserved in both thymocytes, is denoted in purple.

Fig 2. A) Scatter plot for all DN3-called interactions of difference in interaction strength (expressed as log2 (DN3 interaction/DP interaction), after quantile normalization (see *4.3*)) against difference in expression (expressed as log2 (DN3 RPKM/DP RPKM); see *5.3*). The interactions are classed according to these thresholds: A – interaction difference > 1, expression difference > 1; B – interaction difference < 1, expression difference > 1; C – interaction difference > 1, expression difference < -1; D – interaction difference < 1, expression difference < -1; E – interaction difference < 1, abs(expression difference) < 0.5; F– interaction difference > 1, abs(expression difference) < 0.5. Classes E and F can further be split into "silent" or "active" genes, based on their gene group classification (see Fig 1B,C; group 0 = "silent", all others = "active"). C-H) Selected CHi-C and ChIP-seq plots, exactly as for Fig 1F,G. Called interactions are denoted as stripes (blue DN3-specific, red DP-specific, purple present in both thymocyte types).

Fig 3. A) A collection of "strong" and "weak" P5424 enhancers is taken directly from STARR-seq data (Vanhille et al., 2015), and those that intersect with DN3-called interacting regions which fall within interaction classes A-F are extracted. The relative proportions of strong and weak STARR-seq hits falling within each interaction class is plotted. B) Luciferase enhancer activities are compared between P5424 and mES cells by two-tailed t-tests, with Benjamini-Hochberg multiple testing correction. (* $p < 0.05$; ** $p < 0.005$; *** $p < 0.001$). C-E) CHi-C, ChIP-seq and 4C-seq profiles for *Myc* plotted exactly as for Fig 1F-H.

Fig 4. A) and C) Luciferase silencer activities for putative silencers are compared with a cell type-matched "neutral" sequence (different for each cell type) by two-tailed t-tests with Benjamini-Hochberg multiple testing correction (* $p < 0.05$; ** $p < 0.005$; *** $p < 0.001$). Note that whereas the size of the tested inserts are different between A) (2 kb) and C) (< 600 bp), the neutral sequence comparisons are with the same data from a 2 kb insert. B) DN3-interacting regions containing a CTCF ChIP-seq and a facing motif (see explanation in *Promoter Capture analyses*) are scored as to whether or not they have a particular repetitive element within 500 bp of the CTCF motif (repetitive elements directly taken from Repeat Masker database on UCSC). The proportions are plotted for interactions with genes of different expression classes (see Fig 1B,C). P-values are given for Fisher's exact tests comparing the class 0 gene interactions (i.e. inactive) with all the rest (i.e. active).

Fig S1. A-F) Calculated exactly as DN3 equivalent analyses in Fig 1. G,H) Exactly as Fig 1D, E, but specifically for "conserved" CTCF sites (i.e. the intersection of the CTCF ChIP-seq peaks taken for DN3, DP and mES cells), and adding the class of "conserved" interactions (the intersection of PromoMaxima-called interactions for DN3, DP and mES cells).

Figs S2-3. Exactly as for Fig 1F-H for selected viewpoints.

Fig S4. Exactly as for Fig 2A, for different pairwise comparisons of called interactions (the interactions used in the scatter plot are always those called in the numerator cell type).

TADs caller benchmarking,

Fig 1. Heatmaps were produced as *3.1*. The parameters used for calling TADs for each tool are: TADbit (default parameters, identify centromeric regions set to TRUE), Insulation score (insulation square=50 kb, delta span =20 kb), Armatus (gamma-max=0.05, all other

parameters at default setting), Arrowhead (-K= none, -r 5kb). The input for each tool was converted from the original normalized P2 matrices (see *3.1*) using customized scripts.

Transcription directly remodels a small subset of topologically associated domains

Fig 1. A) TAD heat maps were constructed (see *3.3*), and plotted with Arrowhead-called borders (see *3.2*) and ChIP-seq tracks (see *5.2*). B) TAD borders were called as 5 kb bins by Arrowhead (see *3.2*) for the three different cell types and were compared in Venn diagrams. Borders at the extremities of the captured regions were discarded, and TADs were considered conserved if the called 5 kb bin was an exact match in the tested cell types, using the intersect function of bedtools. C) Percentages of "stable" (conserved in DN3 and DP) and "dynamic" (in DN3 but not DP, or in DP but not DN3) borders containing a CTCF ChIP-seq peak (see *5.2*) in DN3 or DP.

Figs 2, 3 and 5. A) Exactly as Fig 1A, for different regions. B) The log2-ratios of DN3/DP CHi-C normalized interaction strengths are plotted (see *3.4*).

Fig 4A. As Fig 1A, but with two cell types/conditions shown in a comparative view by reflecting the second condition along the horizontal. TAD borders called for each separate condition are plotted accordingly.

Fig S1. As Fig 4A, but second condition is actually a heat map derived from the Hi-C data before TAD capture (same visualisation and colour scheme used as *3.3*).

Fig S2. As all other TAD capture matrices.

Fig S3. IGV browser view of quantile normalised (see *5.2*) CTCF ChIP-seq profiles spanning the *Bcl6* locus.

Fig S4. As previous TAD capture matrices and log2-ratio plots.

Abbreviations

3D: 3 dimensions

DNA: Deoxyribonucleic acid

TAD: topological associated domain

HP1: heterochromatin protein 1

LAD: lamina associated domain

RNA: ribonucleic acid

rRNA: ribosomal RNA

mRNA: messenger RNA

tRNA: transfer RNA

PolI: RNA polymerase I

PolII: RNA polymerase II

NAD: Nicotinamide adénine dinucléotide

ES: embryonic stem cells

3C: chromosome conformation capture

Bp: baise pair

SV40: Simien Virus 40

STARR-seq: self-transcribing active regulatory region sequencing

H3K4me1: Histone 3 lysine 4 monomethyaltion

H3K4me2: Histone 3 lysine 4 dimethylation

H3K4me3: Histone 3 lysine 4 trimethylation

H3K27ac: Histone 3 lysine 27 acethylation

H3K27me3: Histone 3 lysine 27 trimethylation

CTCF: CCTC binding factor

TNF-α: Tumor nucreosis factor alpha

Hi-C: High throughput 3C

FISH: fluorescent in situ hybridization

4C: circular 3C

5C: carbon copy 3C

SIM: Super resolution microscopy

TALE: Transcription activator-like effector

CRISPR: clustered regularly interspaced short palindromic repeats

(q)PCR: quantitative polymerase chain reaction

UMI: unique molecular identifiers

ChIA-PET: Chromatin Interaction Analysis by Paired-End Tag Sequencing

HiChIP:

CHi-C: Capture Hi-C

TCR: T cell receptor

ETP: early Thymocyte progentior

DN(1-4): double negative cells

DP: double positive

MHC: major histocompatibility complex

ICE: Iterative Clique enumeration

ROC: receiver operator characteristics

LINE: Long INterspersed Elements

SINE: Short INterspersed Elements

LTR : Long terminal repeats

T-ALL : T acute lymphoblastic leukemia

FACS: Fluorescence activated cell sorting

GFP: Green fluorescent protein

Bibliography

Alipour, E., & Marko, J. F. (2012). Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Research*, *40*(22), 11202–11212.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., & 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.

Anders, A. S. (2014). Package "DESeq."

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., … Duboule, D. (2013). A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science (New York, N.Y.)*, *340*(6137), 1234167.

Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, L. M., Rath, M., & Stark, A. (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science*, *339*(6123), 1074–1077.

Aymard, F., Aguirrebengoa, M., Guillou, E., Javierre, B. M., Bugler, B., Arnould, C., … Legube, G. (2017). Genome-wide mapping of long-range contacts unveils clustering of DNA double-strand breaks at damaged active genes. *Nature Structural & Molecular Biology*, *24*(4), 353–361.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., … Cavalli, G. (2011). Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell*, *144*(2), 214–226.

Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., … Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, *441*(7089), 87–90.

Beliveau, B. J., Boettiger, A. N., Avendaño, M. S., Jungmann, R., McCole, R. B., Joyce, E. F., … Wu, C. (2015). Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. *Nature Communications*, *6*, 7147.

Bell, J. J., & Bhandoola, A. (2008). The earliest thymic progenitors for T cells possess myeloid lineage potential. *Nature*, *452*(7188), 764–767.

Ben Zouari, Yousra Molitor, Anne Sexton, T. (2017). *Sailing the Hi-C's: Benefits and Remaining Challenges in Mapping Chromatin Interaction*.

Berger, S. L. (2007). The complex language of chromatin regulation during transcription. *Nature*, *447*(7143), 407–412.

Bibel, M., Richter, J., Lacroix, E., & Barde, Y.-A. (2007). Generation of a defined and uniform population of CNS progenitors and neurons from mouse embryonic stem cells. *Nature Protocols*, *2*(5), 1034–1043.

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., … Cavalli, G. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, *171*(3), 557–572.e24.

Boyle, S., Gilchrist, S., Bridger, J. M., Mahy, N. L., Ellis, J. A., & Bickmore, W. A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human Molecular Genetics*, *10*(3), 211–219. Retrieved from

Buer, J., Aifantis, I., DiSanto, J. P., Fehling, H. J., & von Boehmer, H. (1997). Role of different T cell receptors in the development of pre-T cells. *The Journal of Experimental Medicine*, *185*(9), 1541–1547.

Cairns, J., Freire-Pritchett, P., Wingett, S. W., Várnai, C., Dimond, A., Plagnol, V., … Spivakov, M. (2016). CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biology*, *17*(1), 127.

Capelson, M., Liang, Y., Schulte, R., Mair, W., Wagner, U., & Hetzer, M. W. (2010). Chromatin-bound nuclear pore components regulate gene expression in higher eukaryotes. *Cell*, *140*(3), 372–383.

Cavalli, G., & Misteli, T. (2013). Functional implications of genome topology. *Nature Structural & Molecular Biology*, *20*(3), 290–299.

Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G.-W., … Huang, B. (2013). Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell*, *155*(7), 1479–1491.

Chen, L., Carico, Z., Shih, H.-Y., & Krangel, M. S. (2015). A discrete chromatin loop in the mouse Tcra-Tcrd locus shapes the TCRδ and TCRα repertoires. *Nature Immunology*, *16*(10), 1085–1093.

Chuang, C.-H., Carpenter, A. E., Fuchsova, B., Johnson, T., de Lanerolle, P., & Belmont, A. S. (2006). Long-range directional movement of an interphase chromosome site. *Current Biology : CB*, *16*(8), 825–831.

Cisse, I. I., Izeddin, I., Causse, S. Z., Boudarene, L., Senecal, A., Muresan, L., … Darzacq, X. (2013). Real-Time Dynamics of RNA Polymerase II Clustering in Live Human Cells. *Science*, *341*(6146), 664–667.

Clowney, E. J., LeGros, M. A., Mosley, C. P., Clowney, F. G., Markenskoff-Papadimitriou,

E. C., Myllys, M., … Lomvardas, S. (2012). Nuclear aggregation of olfactory receptor genes governs their monogenic expression. *Cell*, *151*(4), 724–737.

Conic, S., Desplancq, D., Ferrand, A., Fischer, V., Heyer, V., Reina San Martin, B., … Tora, L. (2018). Imaging of native transcription factors and histone phosphorylation at high resolution in live cells. *The Journal of Cell Biology*, jcb.201709153.

Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R., & Mozziconacci, J. (2012). Normalization of a chromosomal contact map. *BMC Genomics*, *13*(1), 436.

Court, F., Miro, J., Braem, C., Lelay-Taha, M.-N., Brisebarre, A., Atger, F., … Forné, T. (2011). Modulated contact frequencies at gene-rich loci support a statistical helix model for mammalian chromatin organization. *Genome Biology*, *12*(5), R42.

Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., … Meyer, B. J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, *523*(7559), 240–244.

Cremer, T., & Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, *2*(4), 292–301.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., … Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(50), 21931–21936.

Dahm, R. (2005). Friedrich Miescher and the discovery of DNA. *Developmental Biology*, *278*(2), 274–288.

Dalvai, M., Fleury, L., Bellucci, L., Kocanova, S., & Bystricky, K. (2013). TIP48/Reptin and H2A.Z Requirement for Initiating Chromatin Remodeling in Estrogen-Activated Transcription. *PLoS Genetics*, *9*(4), e1003387.

David-Fung, E.-S., Butler, R., Buzi, G., Yui, M. A., Diamond, R. A., Anderson, M. K., … Rothenberg, E. V. (2009). Transcription factor expression dynamics of early T-lymphocyte specification and commitment. *Developmental Biology*, *325*(2), 444–467.

de Bruijn, M. F., & Speck, N. A. (2004). Core-binding factors in hematopoiesis and immune function. *Oncogene*, *23*(24), 4238–4248.

de Wit, E., Bouwman, B. A. M., Zhu, Y., Klous, P., Splinter, E., Verstegen, M. J. A. M., … de Laat, W. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature*, *501*(7466), 227–231.

de Wit, E., Vos, E. S. M., Holwerda, S. J. B., Valdes-Quezada, C., Verstegen, M. J. A. M.,

Teunissen, H., … de Laat, W. (2015a). CTCF Binding Polarity Determines Chromatin Looping. *Molecular Cell*, *60*(4), 676–684.

de Wit, E., Vos, E. S. M., Holwerda, S. J. B., Valdes-Quezada, C., Verstegen, M. J. A. M., Teunissen, H., … de Laat, W. (2015b). CTCF Binding Polarity Determines Chromatin Looping. *Molecular Cell*, *60*(4), 676–684.

Dekker, J. (2006). The three "C" s of chromosome conformation capture: controls, controls, controls. *Nature Methods*, *3*(1), 17–21.

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science (New York, N.Y.)*, *295*(5558), 1306–1311.

Deng, W., Rupon, J. W., Krivega, I., Breda, L., Motta, I., Jahn, K. S., … Blobel, G. A. (2014). Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell*, *158*(4), 849–860.

Denholtz, M., Bonora, G., Chronis, C., Splinter, E., de Laat, W., Ernst, J., … Plath, K. (2013). Long-Range Chromatin Contacts in Embryonic Stem Cells Reveal a Role for Pluripotency Factors and Polycomb Proteins in Genome Organization. *Cell Stem Cell*, *13*(5), 602–616.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., … 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., … Ren, B. (2012a). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380.

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

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., … Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research*, *16*(10), 1299–1309.

Drissen, R., Palstra, R.-J., Gillemans, N., Splinter, E., Grosveld, F., Philipsen, S., & de Laat, W. (2004). The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes & Development*, *18*(20), 2485–2490.

Dryden, N. H., Broome, L. R., Dudbridge, F., Johnson, N., Orr, N., Schoenfelder, S., …

Fletcher, O. (2014). Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Research*, *24*(11), 1854–1868.

Dubarry, M., Loïodice, I., Chen, C. L., Thermes, C., & Taddei, A. (2011). Tight protein-DNA interactions favor gene silencing. *Genes & Development*, *25*(13), 1365–1370.

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*, *3*(1), 95–98.

Egawa, T., & Littman, D. R. (2011). Transcription factor AP4 modulates reversible and epigenetic silencing of the Cd4 gene. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(36), 14873–14878.

Fabre, P. J., Benke, A., Joye, E., Nguyen Huynh, T. H., Manley, S., & Duboule, D. (2015). Nanoscale spatial organization of the HoxD gene cluster in distinct transcriptional states. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(45), 13964–13969.

Fabre, P. J., Benke, A., Manley, S., & Duboule, D. (2015). Visualizing the HoxD Gene Cluster at the Nanoscale Level. *Cold Spring Harbor Symposia on Quantitative Biology*, *80*, 9–16.

Fanucchi, S., Shibayama, Y., Burd, S., Weinberg, M. S., & Mhlanga, M. M. (2013). Chromosomal Contact Permits Transcription between Coregulated Genes. *Cell*, *155*(3), 606–620.

Filippova, D., Patro, R., Duggal, G., & Kingsford, C. (2014a). Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology*, *9*(1), 14.

Filippova, D., Patro, R., Duggal, G., & Kingsford, C. (2014b). Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology : AMB*, *9*(1), 14.

Finlan, L. E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., … Bickmore, W. A. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genetics*, *4*(3), e1000039.

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

Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., … Mundlos, S. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, *538*(7624), 265–269.

Freire-Pritchett, P., Schoenfelder, S., Várnai, C., Wingett, S. W., Cairns, J., Collier, A. J., …

Spivakov, M. (2017). Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *eLife*, *6*.

Friedli, M., & Trono, D. (2015). The Developmental Control of Transposable Elements and the Evolution of Higher Species. *Annual Review of Cell and Developmental Biology*, *31*(1), 429–451.

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

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. Bin, … Ruan, Y. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, *462*(7269), 58–64.

Gall, J. G., & Pardue, M. L. (1969). Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proceedings of the National Academy of Sciences of the United States of America*, *63*(2), 378–383.

Germier, T., Kocanova, S., Walther, N., Bancaud, A., Shaban, H. A., Sellou, H., … Bystricky, K. (2017). Real-Time Imaging of a Single Gene Reveals Transcription-Initiated Local Confinement. *Biophysical Journal*, *113*(7), 1383–1394.

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

Gibcus, J. H., Samejima, K., Goloborodko, A., Samejima, I., Naumova, N., Nuebler, J., … Dekker, J. (2018). A pathway for mitotic chromosome formation. *Science*, *359*(6376), eaao6135.

Giorgetti, L., Galupa, R., Nora, E. P., Piolot, T., Lam, F., Dekker, J., … Heard, E. (2014). Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, *157*(4), 950–963.

Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V, Martin, A. R., … Snyder, M. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, *162*(5), 1051–1065.

Gu, B., Swigut, T., Spencley, A., Bauer, M. R., Chung, M., Meyer, T., & Wysocka, J. (2018). Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science*, *359*(6379), 1050–1055.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D. U., … Wu, Q. (2015). CRISPR

Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*, *162*(4), 900–910.

Haarhuis, J. H. I., van der Weide, R. H., Blomen, V. A., Yáñez-Cuna, J. O., Amendola, M., van Ruiten, M. S., … Rowland, B. D. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell*, *169*(4), 693–707.e14.

Hakim, O., Sung, M.-H., Voss, T. C., Splinter, E., John, S., Sabo, P. J., … Hager, G. L. (2011). Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements. *Genome Research*, *21*(5), 697–706.

Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., … Wei, C.-L. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature Genetics*, *43*(7), 630–638.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., … Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, *459*(7243), 108–112.

Herranz, D., Ambesi-Impiombato, A., Palomero, T., Schnell, S. A., Belver, L., Wendorff, A. A., … Ferrando, A. A. (2014). A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. *Nature Medicine*, *20*(10), 1130–1137.

Hu, G., Cui, K., Fang, D., Hirose, S., Wang, X., Wangsa, D., … Zhao, K. (2018). Transformation of Accessible Chromatin and 3D Nucleome Underlies Lineage Commitment of Early T Cells. *Immunity*, *48*(2), 227–242.e8.

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., & Liu, J. S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics (Oxford, England)*, *28*(23), 3131–3133.

Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., … Higgs, D. R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics*, *46*(2), 205–212.

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

Isoda, T., Moore, A. J., He, Z., Chandra, V., Aida, M., Denholtz, M., … 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.e18.

Jackson, D. A., & Cook, P. R. (1985). Transcription occurs at a nucleoskeleton. *The EMBO Journal*, *4*(4), 919–925.

Jäger, R., Migliorini, G., Henrion, M., Kandaswamy, R., Speedy, H. E., Heindl, A., … Houlston, R. S. (2015). Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature Communications*, *6*, 6178.

Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., … Flicek, P. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, *167*(5), 1369–1384.e19.

Jiang, H., & Peterlin, B. M. (2008). Differential Chromatin Looping Regulates CD4 Expression in Immature Thymocytes. *Molecular and Cellular Biology*, *28*(3), 907–912.

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., … Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, *503*(7475), 290–294.

Jing, H., Vakoc, C. R., Ying, L., Mandat, S., Wang, H., Zheng, X., & Blobel, G. A. (2008). Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. *Molecular Cell*, *29*(2), 232–242.

Joshi, O., Wang, S.-Y., Kuznetsova, T., Atlasi, Y., Peng, T., Fabre, P. J., … Stunnenberg, H. G. (2015). Dynamic Reorganization of Extremely Long-Range Promoter-Promoter Interactions between Two States of Pluripotency. *Cell Stem Cell*, *17*(6), 748–757.

Kadesch, T., Zervos, P., & Ruezinsky, D. (1986). Functional analysis of the murine IgH enhancer: evidence for negative control of cell-type specificity. *Nucleic Acids Research*, *14*(20), 8209–8221.

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., … Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, *467*(7314), 430–435.

Kanhere, A., Hertweck, A., Bhatia, U., Gökmen, M. R., Perucha, E., Jackson, I., … Jenner, R. G. (2012). T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements. *Nature Communications*, *3*(1), 1268.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., … Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187.

Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T. K., Zacarias-Cabeza, J., … Andrau, J.-C. (2011). Transcription initiation platforms and GTF recruitment at tissue-specific

enhancers and promoters. *Nature Structural & Molecular Biology*, *18*(8), 956–963.

Kolovos, P., van de Werken, H. J., Kepper, N., Zuin, J., Brouwer, R. W., Kockx, C. E., … Knoch, T. A. (2014). Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics & Chromatin*, *7*, 10.

Kosak, S. T., Skok, J. A., Medina, K. L., Riblet, R., Le Beau, M. M., Fisher, A. G., & Singh, H. (2002). Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. *Science (New York, N.Y.)*, *296*(5565), 158–162.

Kurukuti, S., Tiwari, V. K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., … Ohlsson, R. (2006). CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(28), 10684–10689.

Lavelle, C. (2014). Pack, unpack, bend, twist, pull, push: the physical side of gene expression. *Current Opinion in Genetics & Development*, *25*, 74–84.

Le Dily, F., Bau, D., Pohl, a., Vicent, G. P., Serra, F., Soronellas, D., … Beato, M. (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & Development*, *28*(19), 2151–2162.

Le May, N., Fradin, D., Iltis, I., Bougnères, P., & Egly, J.-M. (2012). XPG and XPF endonucleases trigger chromatin looping and DNA demethylation for accurate expression of activated genes. *Molecular Cell*, *47*(4), 622–632.

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., … Ruan, Y. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, *148*(1–2), 84–98.

Li, L., Zhang, J. A., Dose, M., Kueh, H. Y., Mosadeghi, R., Gounari, F., & Rothenberg, E. V. (2013). A far downstream enhancer for murine Bcl11b controls its T-cell specific expression. *Blood*, *122*(6), 902–911.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., … Dekker, J. (2009a). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, *326*(5950), 289–293. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., … Dekker, J. (2009b). Comprehensive mapping of long-range interactions reveals

folding principles of the human genome. *Science (New York, N.Y.)*, *326*(5950), 289–293.

Lowe, C. B., Bejerano, G., & Haussler, D. (2007). Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences*, *104*(19), 8005–8010.

Lucas, J. S., Zhang, Y., Dudko, O. K., & Murre, C. (2014). 3D trajectories adopted by coding and regulatory DNA elements: first-passage times for genomic interactions. *Cell*, *158*(2), 339–352.

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., … Mundlos, S. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, *161*(5), 1012–1025.

Ma, H., Tu, L.-C., Naseri, A., Huisman, M., Zhang, S., Grunwald, D., & Pederson, T. (2016). Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. *Nature Biotechnology*, *34*(5), 528–530.

Mateos-Langerak, J., Bohn, M., de Leeuw, W., Giromus, O., Manders, E. M. M., Verschure, P. J., … Goetze, S. (2009). Spatially confined folding of chromatin in the interphase nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(10), 3812–3817.

May, E., Omilli, F., Ernoult-Lange, M., Zenke, M., & Chambon, P. (1987). The sequence motifs that are involved in SV40 enhancer function also control SV40 late promoter activity. *Nucleic Acids Research*, *15*(6), 2445–2461. Retrieved from

Mendjan, S., Taipale, M., Kind, J., Holz, H., Gebhardt, P., Schelder, M., … Akhtar, A. (2006). Nuclear Pore Components Are Involved in the Transcriptional Regulation of Dosage Compensation in Drosophila. *Molecular Cell*, *21*(6), 811–823.

Meuleman, W., Peric-Hupkes, D., Kind, J., Beaudry, J.-B., Pagie, L., Kellis, M., … van Steensel, B. (2013). Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Research*, *23*(2), 270–280.

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., … Osborne, C. S. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, *47*(6), 598–606.

Miyanari, Y., Ziegler-Birling, C., & Torres-Padilla, M.-E. (2013). Live visualization of chromatin dynamics with fluorescent TALEs. *Nature Structural & Molecular Biology*, *20*(11), 1321–1324.

Morgan, S. L., Mariano, N. C., Bermudez, A., Arruda, N. L., Wu, F., Luo, Y., … Wang, K. C.

(2017). Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping. *Nature Communications*, *8*, 15993.

Muerdter, F., Boryń, Ł. M., Woodfin, A. R., Neumayr, C., Rath, M., Zabidi, M. A., … Stark, A. (2017). Resolving systematic errors in widely used enhancer activity assays in human cells. *Nature Methods*, *15*(2), 141–149.

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

Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., … Chang, H. Y. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nature Genetics*, *49*(11), 1602–1612.

Narendra, V., Bulajić, M., Dekker, J., Mazzoni, E. O., & Reinberg, D. (2016). CTCF-mediated topological boundaries during development foster appropriate gene regulation. *Genes & Development*, *30*(24), 2657–2662.

Narendra, V., Rocha, P. P., An, D., Raviram, R., Skok, J. A., Mazzoni, E. O., & Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science*, *347*(6225), 1017–1021.

Nasmyth, K. (2001). Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annual Review of Genetics*, *35*(1), 673–745.

Nasmyth, K., & Haering, C. H. (2009). Cohesin: Its Roles and Mechanisms. *Annual Review of Genetics*, *43*(1), 525–558.

Németh, A., Conesa, A., Santoyo-Lopez, J., Medina, I., Montaner, D., Péterfia, B., … Längst, G. (2010). Initial genomics of the human nucleolus. *PLoS Genetics*, *6*(3), e1000889.

Ng, S. Y.-M., Yoshida, T., Zhang, J., & Georgopoulos, K. (2009). Genome-wide lineage-specific transcriptional networks underscore Ikaros-dependent lymphoid priming in hematopoietic stem cells. *Immunity*, *30*(4), 493–507.

Noordermeer, D., Leleu, M., Schorderet, P., Joye, E., Chabaud, F., & Duboule, D. (2014). Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci. *eLife*, *3*, e02557.

Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W., & Duboule, D. (2011). The dynamic architecture of Hox gene clusters. *Science (New York, N.Y.)*, *334*(6053), 222–225.

Nora, E. P., Goloborodko, A., Valton, A.-L., Gibcus, J. H., Uebersohn, A., Abdennur, N., …

Bruneau, B. G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, *169*(5), 930–944.e22.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., … Heard, E. (2012a). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381–385.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., … Heard, E. (2012b). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381–385.

Oravecz, A., Apostolov, A., Polak, K., Jost, B., Le Gras, S., Chan, S., & Kastner, P. (2015). Ikaros mediates gene silencing in T cells through Polycomb repressive complex 2. *Nature Communications*, *6*(1), 8823.

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., … Fraser, P. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genetics*, *36*(10), 1065–1071.

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., … Pennacchio, L. A. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, *554*(7691), 239–243.

Palstra, R.-J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., & de Laat, W. (2003). The beta-globin nuclear compartment in development and erythroid differentiation. *Nature Genetics*, *35*(2), 190–194.

Papantonis, A., Kohro, T., Baboo, S., Larkin, J. D., Deng, B., Short, P., … Cook, P. R. (2012). TNFα signals through specialized factories where responsive coding and miRNA genes are transcribed. *The EMBO Journal*, *31*(23), 4404–4414.

Parada, L. A., McQueen, P. G., & Misteli, T. (2004). Tissue-specific spatial organization of genomes. *Genome Biology*, *5*(7), R44.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., … Merkenschlager, M. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, *132*(3), 422–433.

Patel, N. S., Rhinn, M., Semprich, C. I., Halley, P. A., Dollé, P., Bickmore, W. A., & Storey, K. G. (2013). FGF Signalling Regulates Chromatin Organisation during Neural Differentiation via Mechanisms that Can Be Uncoupled from Transcription. *PLoS Genetics*, *9*(7), e1003614.

Pekowska, A., Benoukraf, T., Zacarias-Cabeza, J., Belhocine, M., Koch, F., Holota, H., … Spicuglia, S. (2011). H3K4 tri-methylation provides an epigenetic signature of active enhancers. *The EMBO Journal*, *30*(20), 4198–4210.

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W. M., Solovei, I., Brugman, W., … van Steensel, B. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular Cell*, *38*(4), 603–613.

Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., … Corces, V. G. (2013a). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, *153*(6), 1281–1295.

Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., … Corces, V. G. (2013b). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, *153*(6), 1281–1295.

Phillips, J. E., & Corces, V. G. (2009). CTCF: master weaver of the genome. *Cell*, *137*(7), 1194–1211.

PLOS ONE Staff. (2017). Correction: GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PLOS ONE*, *12*(5), e0177280.

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., … Gilbert, D. M. (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature*, *515*(7527), 402–405.

Ptak, C., Aitchison, J. D., & Wozniak, R. W. (2014). The multifunctional nuclear pore complex: a platform for controlling gene expression. *Current Opinion in Cell Biology*, *28*, 46–53.

Raab, J. R., Chiu, J., Zhu, J., Katzman, S., Kurukuti, S., Wade, P. A., … Kamakaka, R. T. (2012). Human tRNA genes function as chromatin insulators. *The EMBO Journal*, *31*(2), 330–350.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, *470*(7333), 279–283.

Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M. D., Banerjee, B., … Sherwood, R. I. (2016). High-throughput mapping of regulatory DNA. *Nature Biotechnology*, *34*(2), 167–174.

Rao, S. S. P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K.-R., … Aiden, E. L. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell*, *171*(2),

305–320.e24.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014a). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, *159*(7), 1665–1680.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014b). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, *159*(7), 1665–1680.

Reddy, K. L., Zullo, J. M., Bertolino, E., & Singh, H. (2008). Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature*, *452*(7184), 243–247.

Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., … Zhao, K. (2017). CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Molecular Cell*, *67*(6), 1049–1058.e6.

Roadmap Epigenomics Consortium, A., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330.

Robinett, C. C., Straight, A., Li, G., Willhelm, C., Sudlow, G., Murray, A., & Belmont, A. S. (1996). In vivo localization of DNA sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition. *The Journal of Cell Biology*, *135*(6 Pt 2), 1685–1700.

Rodríguez-Navarro, S., Fischer, T., Luo, M.-J., Antúnez, O., Brettschneider, S., Lechner, J., … Hurt, E. (2004). Sus1, a Functional Component of the SAGA Histone Acetylase Complex and the Nuclear Pore-Associated mRNA Export Machinery. *Cell*, *116*(1), 75–86.

Rowley, M. J., Nichols, M. H., Lyu, X., Ando-Kuri, M., Rivera, I. S. M., Hermetz, K., … Corces, V. G. (2017). Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Molecular Cell*, *67*(5), 837–852.e7.

Rubin, A. J., Barajas, B. C., Furlan-Magaril, M., Lopez-Pajares, V., Mumbach, M. R., Howard, I., … Khavari, P. A. (2017). Lineage-specific dynamic and pre-established enhancer–promoter contacts cooperate in terminal differentiation. *Nature Genetics*, *49*(10), 1522–1528.

Saad, H., Gallardo, F., Dalvai, M., Tanguy-le-Gac, N., Lane, D., & Bystricky, K. (2014). DNA dynamics during early double-strand break processing revealed by non-intrusive imaging of living cells. *PLoS Genetics*, *10*(3), e1004187.

Sahlén, P., Abdullayev, I., Ramsköld, D., Matskova, L., Rilakovic, N., Lötstedt, B., … Sandberg, R. (2015). Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biology*, *16*(1), 156.

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., … Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(47), E6456-65.

Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, *489*(7414), 109–113.

Sasaki, T., Nishihara, H., Hirakawa, M., Fujimura, K., Tanaka, M., Kokubo, N., … Okada, N. (2008). Possible involvement of SINEs in mammalian-specific brain formation. *Proceedings of the National Academy of Sciences*, *105*(11), 4220–4225.

Saurin, A. J., Shiels, C., Williamson, J., Satijn, D. P., Otte, A. P., Sheer, D., & Freemont, P. S. (1998). The human polycomb group complex associates with pericentromeric heterochromatin to form a novel nuclear domain. *The Journal of Cell Biology*, *142*(4), 887–898. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9722603

Sawada, S., Scarborough, J. D., Killeen, N., & Littman, D. R. (1994). A lineage-specific transcriptional silencer regulates CD4 gene expression during T lymphocyte development. *Cell*, *77*(6), 917–929.

Schlenner, S. M., & Rodewald, H.-R. (2010). Early T cell development and the pitfalls ofSchlenner, S. M., & Rodewald, H.-R. (2010). Early T cell development and the pitfalls of potential.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., … Fraser, P. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, *25*(4), 582–597.

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N. F., Horton, A., Andrews, S., … Fraser, P. (2010). Preferential associations between co-regulated genes reveal a transcriptional

interactome in erythroid cells. *Nature Genetics*, *42*(1), 53–61.

Schoenfelder, S., Sugar, R., Dimond, A., Javierre, B.-M., Armstrong, H., Mifsud, B., … Elderkin, S. (2015). Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nature Genetics*, *47*(10), 1179–1186.

Schwartzman, O., Mukamel, Z., Oded-Elkayam, N., Olivares-Chauvet, P., Lubling, Y., Landan, G., … Tanay, A. (2016). UMI-4C for quantitative and targeted chromosomal contact profiling. *Nature Methods*, *13*(8), 685–691.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., … Spitz, F. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, *551*(7678), 51–56.

Scripture-Adams, D. D., Damle, S. S., Li, L., Elihu, K. J., Qin, S., Arias, A. M., … Rothenberg, E. V. (2014). GATA-3 dose-dependent checkpoints in early T cell commitment. *Journal of Immunology (Baltimore, Md. : 1950)*, *193*(7), 3470–3491.

Seitan, V. C., Faure, A. J., Zhan, Y., McCord, R. P., Lajoie, B. R., Ing-Simmons, E., … Merkenschlager, M. (2013). Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Research*, *23*(12), 2066–2077.

Serra, F., Baù, D., Filion, G., & Marti-Renom, M. A. (2016). *Structural features of the fly chromatin colors revealed by automatic three-dimensional modeling. bioRxiv*. Cold Spring Harbor Labs Journals.

Sexton, T., & Cavalli, G. (2015). Review The Role of Chromosome Domains in Shaping the Functional Genome. *Cell*, *160*(6), 1049–1059. https://doi.org/10.1016/j.cell.2015.02.040

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., … Cavalli, G. (2012a). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, *148*(3), 458–472.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., … Cavalli, G. (2012b). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, *148*(3), 458–472.

Shachar, S., Voss, T. C., Pegoraro, G., Sciascia, N., & Misteli, T. (2015). Identification of Gene Positioning Factors Using High-Throughput Imaging Mapping. *Cell*, *162*(4), 911–923.

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., … Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*, *488*(7409), 116–120.

Siersbæk, R., Madsen, J. G. S., Javierre, B. M., Nielsen, R., Bagge, E. K., Cairns, J., … Mandrup, S. (2017). Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation. *Molecular Cell*, *66*(3), 420–435.e5.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., … de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics*, *38*(11), 1348–1354.

Sofueva, S., Yaffe, E., Chan, W.-C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., … Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. *The EMBO Journal*, *32*(24), 3119–3129.

Solovei, I., Kreysing, M., Lanctôt, C., Kösem, S., Peichl, L., Cremer, T., … Joffe, B. (2009). Nuclear Architecture of Rod Photoreceptor Cells Adapts to Vision in Mammalian Evolution. *Cell*, *137*(2), 356–368.

Somech, R., Shaklai, S., Geller, O., Amariglio, N., Simon, A. J., Rechavi, G., & Gal-Yam, E. N. (2005). The nuclear-envelope protein and transcriptional repressor LAP2 interacts with HDAC3 at the nuclear periphery, and induces histone H4 deacetylation. *Journal of Cell Science*, *118*(17), 4017–4025.

Soshnikova, N., & Duboule, D. (2009). Epigenetic temporal control of mouse Hox genes in vivo. *Science (New York, N.Y.)*, *324*(5932), 1320–1323.

Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., … de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & Development*, *20*(17), 2349–2354.

Swedlow, J. R., & Hirano, T. (2003). The Making of the Mitotic Chromosome: Modern Insights into Classical Questions. *Molecular Cell*, *11*(3), 557–569.

Sydor, A. M., Czymmek, K. J., Puchner, E. M., & Mennella, V. (2015). Super-Resolution Microscopy: From Single Molecules to Supramolecular Assemblies. *Trends in Cell Biology*, *25*(12), 730–748.

Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., … Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Research*, *24*(3), 390–400.

Talbot, D., Collis, P., Antoniou, M., Vidal, M., Grosveld, F., & Greaves, D. R. (1989). A dominant control region from the human β-globin locus conferring integration site-independent gene expression. *Nature*, *338*(6213), 352–355.

Tan-Wong, S. M., Zaugg, J. B., Camblong, J., Xu, Z., Zhang, D. W., Mischo, H. E., …

Proudfoot, N. J. (2012). Gene loops enhance transcriptional directionality. *Science (New York, N.Y.)*, *338*(6107), 671–675.

Thompson, P. K., & Zúñiga-Pflücker, J. C. (2011). On becoming a T cell, a convergence of factors kick it up a Notch along the way. *Seminars in Immunology*, *23*(5), 350–359.

Thomson, I., Gilchrist, S., Bickmore, W. A., & Chubb, J. R. (2004). The radial positioning of chromatin is not inherited through mitosis but is established de novo in early G1. *Current Biology : CB*, *14*(2), 166–172.

Tinsley, K. W., Hong, C., Luckey, M. A., Park, J.-Y., Kim, G. Y., Yoon, H. -w., … Park, J.-H. (2013). Ikaros is required to survive positive selection and to maintain clonal diversity during T-cell development in the thymus. *Blood*, *122*(14), 2358–2368.

Vakoc, C. R., Letting, D. L., Gheldof, N., Sawado, T., Bender, M. A., Groudine, M., … Blobel, G. A. (2005). Proximity among Distant Regulatory Elements at the β-Globin Locus Requires GATA-1 and FOG-1. *Molecular Cell*, *17*(3), 453–462.

van de Corput, M. P. C., de Boer, E., Knoch, T. A., van Cappellen, W. A., Quintanilla, A., Ferrand, L., & Grosveld, F. G. (2012). Super-resolution imaging reveals three-dimensional folding dynamics of the β-globin locus upon gene activation. *Journal of Cell Science*, *125*(Pt 19), 4630–4639.

van de Werken, H. J. G., Landan, G., Holwerda, S. J. B., Hoichman, M., Klous, P., Chachik, R., … de Laat, W. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods*, *9*(10), 969–972.

van Koningsbruggen, S., Gierlinski, M., Schofield, P., Martin, D., Barton, G. J., Ariyurek, Y., … Lamond, A. I. (2010). High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. *Molecular Biology of the Cell*, *21*(21), 3735–3748.

Vanhille, L., Griffon, A., Maqbool, M. A., Zacarias-Cabeza, J., Dao, L. T. M., Fernandez, N., … Spicuglia, S. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nature Communications*, *6*(1), 6905.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., & Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Reports*, *10*(8), 1297–1309.

Vietri Rudan, M., Hadjur, S., & Sexton, T. (2017). Detecting Spatial Chromatin Organization by Chromosome Conformation Capture II: Genome-Wide Profiling by Hi-C. *Methods in Molecular Biology (Clifton, N.J.)*, *1589*, 47–74.

Wang, S., Su, J.-H., Beliveau, B. J., Bintu, B., Moffitt, J. R., Wu, C., & Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. *Science (New York, N.Y.)*, *353*(6299), 598–602.

West, A. G., & Fraser, P. (2005). Remote control of gene transcription. *Human Molecular Genetics*, *14 Spec No 1*(suppl_1), R101-11.

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., & Andrews, S. (2015). HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research*, *4*, 1310.

Wutz, G., Varnai, C., Nagasaka, K., Cisneros, D. A., Stocsits, R., Tang, W., … Peters, J.-M. (2017). CTCF, WAPL and PDS5 proteins control the formation of TADs and loops by cohesin. *bioRxiv*, 177444.

Yaffe, E., & Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, *43*(11), 1059–1065.

Ye, Q., Callebaut, I., Pezhman, A., Courvalin, J. C., & Worman, H. J. (1997). Domain-specific interactions of human HP1-type chromodomain proteins and inner nuclear membrane protein LBR. *The Journal of Biological Chemistry*, *272*(23), 14983–14989.

Zhang, H., Jiao, W., Sun, L., Fan, J., Chen, M., Wang, H., … Hu, J.-F. (2013). Intrachromosomal Looping Is Required for Activation of Endogenous Pluripotency Genes during Reprogramming. *Cell Stem Cell*, *13*(1), 30–35.

Zhang, J. A., Mortazavi, A., Williams, B. A., Wold, B. J., & Rothenberg, E. V. (2012). Dynamic Transformations of Genome-wide Epigenetic Marking and Transcriptional Control Establish T Cell Identity. *Cell*, *149*(2), 467–482.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., … Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, *38*(11), 1341–1347.

Zhou, X., Lowdon, R. F., Li, D., Lawson, H. A., Madden, P. A. F., Costello, J. F., & Wang, T. (2013). Exploring long-range genome interactions using the WashU Epigenome Browser. *Nature Methods*, *10*(5), 375–376.

Zuin, J., Dixon, J. R., van der Reijden, M. I. J. A., Ye, Z., Kolovos, P., Brouwer, R. W. W., … Wendt, K. S. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences of the*

United States of America, *111*(3), 996–1001.

Yousra BEN ZOUARI

The functional and spatial organization of chromatin during Thymocyte development

Résumé

Malgré les vastes études démontrant le rôle de la conformation génomique dans le contrôle transcriptionnel, de nombreuses questions restent en suspens, et en particulier, comment ces structures chromatiniennes sont formées et maintenues. Pour mieux comprendre les liens entre l'état de la chromatine au niveau des éléments régulateurs, la topologie de la chromatine et la régulation de la transcription, nous utilisons la technique CHi-C basée sur la technologie de capture de la conformation chromosomique (3C). En utilisant deux stratégies de capture ciblant deux différentes structure chromatiniennes (les boucles chromatiniennes et les domaines topologiques), nous avons pu décrypter la structure chromatinienne associée à la différenciation des thymocytes et mettre en évidence des mécanismes de contrôle transcriptionnel de certains gènes. Les expériences futures de l'équipe vont consister à examiner les facteurs (hors transcription) qui peuvent influencer l'architecture de la chromatine, comme la liaison différentielle des CTCF, et comment ces facteurs peuvent être coordonnés par le contrôle de transcription.

Summary

Chromosome folding takes place at different hierarchical levels, with various topologies correlated with control of gene expression. Despite the large number of recent studies describing chromatin topologies and their correlations with gene activity, many questions remain, in particular how these topologies are formed and maintained. To understand better the link between epigenetic marks, chromatin topology and transcriptional control, we use CHi-C technique based on the chromosome conformation capture (3C) method. By using two capture strategies targeting two different chromatin structures (chromatin loops and topological domains), we have been able to decipher the chromatin structure associated with thymocyte differentiation and to highlight mechanisms for the transcriptional control of certain genes. Future experiments of the lab will examine mechanisms other than transcription which may influence chromatin architecture, such as differential binding of CTCF, and how these may interplay with transcriptional control and chromatin architecture.

Key words: 3D genome architecture, CHi-C, Hi-C, thymocyte differentiation, transcription factors, enhancers, transcription, epigenetic, chromatin loops, TADs