Chromatin 3D modelling from sparse 3C-based datasets

Julen Mendieta Esteban

TESI DOCTORAL UPF / 2020

Thesis supervisors:

Dr. Marc A. Marti-Renom

Dr. Irene Farabella

Department of Experimental and Health Sciences

STRUCTURAL GENOMICS

CENTRE NACIONAL D'ANÀLISI GENÒMICA

GENE REGULATION, STEM CELLS AND CANCER

CENTRE FOR GENOMIC REGULATION

Dedication and acknowledgements

Txapelaz gain, bizipen guztiekin abiatuko naiz irekitzear zaidan mundura. Eskerrik asko guztiagatik ama, aita, Oiana, familia eta lagunak.

Abstract

Genome spatial organisation and transcriptional activity are tightly coordinated to ensure the correct function of the cell. Thus, proper understanding of the chromatin organisation is needed to deepen into the processes regulating the activity of specific loci of interest. In this matter, Chromatin Conformation Capture (3C)-based technologies have helped to increase the understanding of the genomic interaction landscape. Particularly, sparse 3C technologies, like promoter capture Hi-C (pcHi-C), have focused on specific interactions of interest to unveil the interaction landscape associated with functional elements, like promoters. However, to properly characterize the sparse interaction profiles of pcHi-C, it is important to contextualize these interactions in a 3D perspective. Hence, in this thesis, we have developed a tool for the 3D modelling and analysis of sparse 3C-based datasets like pcHi-C, and we have probed its utility to help interpreting the regulatory architecture surrounding genes associated with cell-type or tissue-specific activity.

Resumen

La organización espacial del genoma y la actividad transcripcional están estrechamente coordinadas para garantizar el correcto funcionamiento de la célula. Por lo tanto, se necesita una comprensión adecuada de la organización de la cromatina para profundizar en los procesos que regulan la actividad de *loci* de interés. Tecnologías basadas en la captura de conformación de cromatina (3C) han facilitado la comprensión de la arquitectura genómica. Particularmente, las tecnologías 3C *sparse*, como *promoter capture Hi-C* (pcHi-C), se han centrado en interacciones específicas de interés para desvelar el panorama de interacción asociado con elementos funcionales como los promotores. Sin embargo, para comprender adecuadamente los perfiles *sparse* de interacción de pcHi-C, es importante contextualizar la perspectiva 3D que subyace a estas interacciones. En esta tesis, hemos desarrollado una herramienta para el modelado y análisis 3D de datos *sparse* derivados de 3C como pcHi-C, y hemos probado su utilidad en la comprensión de la arquitectura reguladora de genes asociados con una actividad específica del tipo celular o tejido.

Preface

The genomic content is encoded in chains of instructions needed for the proper proliferation and function of the cell. In Eukaryotes it is enclosed inside of the cell nucleus and undergoes various steps of compaction and compartmentalisation that ensure its functionality. Starting from the DNA double helix, the first step of compaction involves the wrapping of the DNA around the histone octamer to form the basic unit of the chromatin, the nucleosome complex. Nucleosomes can further arrange in groups of variable density, conforming fibres that range between 11 and 30 nm width. The width of this fibre has an effect on how accessible the chromatin is to regulatory proteins associated with promoter and enhancer activity, among others. Interestingly, the genome tends to distribute in the cell nucleus by aggregation of accessible and non-accessible areas into compartments. These compartments, likewise segregate at different scales into high-frequency interacting areas, defining self-interacting domains or TADs and long-range chromatin loops and interactions.

New molecular biology methods based on Chromatin Conformation Capture (3C), together with microscopy imaging-based technologies, have helped to shed light on the forces driving chromatin architecture and dynamics from the whole genome to the locus-specific scale. In this way, they have also shed light into the genome organisation changes associated with cell disease and malfunction. Some 3C-based technologies have focused on the analysis of genome-wide interactions involving specific dispersed loci in the genome with important functional roles, as promoters in the case of promoter capture Hi-C (pcHi-C). However, due to the difficulties associated with the analysis of sparse datasets, and the novelty of the technology, there are few tools available for the analysis of these datasets and none of them takes into account their original 3D context.

This thesis is composed of multiple chapters. In the introduction, we review the processes driving the genome organisation, and their importance for the proper function of the cell. We also review methods for the analysis of this architecture by molecular-biology-based and imaging-based technologies, and different strategies for the 3D modelling of the chromatin. The core of the thesis, in

chapters 1 and 2 presents the results obtained in the two main publications of the candidate, whereas annexe 1 shows an application of the developed method orthogonal to the scope of the thesis. In chapter 1, we present a chromatin 3D modelling approach, focused on the normalisation, 3D modelling, and further analysis of sparse 3C-based datasets like pcHi-C. In chapter 2, we show an application of the method to analyse enhancer 3D hubs in regions containing key regulatory elements associated with type 2 diabetes. In annexe 1, we present an application of the method on a dense 3C-based dataset, specifically in Hi-C, to analyse the differential organisation of two loci before and after CTCF depletion. Finally, a conclusion is added to highlight the main contributions of this thesis.

Objectives

The main objective of this thesis is to provide a reliable tool for the modelling and further analysis of sparse Chromatin Conformation Capture (3C)-based datasets. Specifically, we focused on the reconstruction and analysis of promoter capture Hi-C (pcHi-C) datasets, and subdivided the work into two projects:

1. We developed and tested a tool for the modelling of pcHi-C datasets, paying particular attention to the reliability of the obtained models and the limitations associated with the amount of available data.

    - Additionally, we designed new methods for the differential organisation analysis of the chromatin.

2. We applied the new tool for the analysis of the chromatin organisation in loci associated with the development of type 2 diabetes.

Table of contents

Pag.

List of figures

Pag.

INTRODUCTION

1 DNA structure and organisation

The cell nucleus encloses, and at the same time protects, the instructions booklet for the formation, maintenance, and function of eukaryotic life: the DNA. DNA's information is encoded in chains of molecules named nucleotides, which are classified into four types according to the nitrogenous base that conform them: Cytosine (C), Guanine (G), Adenine (A), or Thymine (T). These nitrogenous bases are complementary to each other in pairs. Specifically, A is complementary to T, and C to G. This allows the covalent joining and stabilisation of two polynucleotide chains with a complementary sequence, thus containing the same biological information, to form a coiling structure named as the DNA double helix (**Figure 1.1**). The redundancy of information facilitates the accessibility to the DNA content and ensures the recovery of damaged DNA strands by using its complementary template.

Around 1% of the DNA sequence in humans encodes information for the transcription and subsequent translation of RNA into proteins, which will be actively involved in most of the chemical processes of the cell (*Bernstein, Birney et al. 2012*). At the same time, some proteins are involved in regulatory processes by their interactions with the non-coding DNA, which represents the remaining 99% of the DNA sequence. Non-coding DNA contains regions with regulatory function like: i) non-coding RNAs (transcribed RNA molecules that although not translated into proteins, are involved in many steps of gene regulation, transcription, and translation for instance) (*Zhang, Wu et al. 2019*); ii) enhancers (target regions for protein binding that can modulate the transcription of a particular gene or set of genes); and iii) promoters (protein binding regions associated with the transcription initiation of the nearest gene in the DNA sequence) (*Zabidi and Stark 2016*). Since these regions are dispersed through the DNA and do not necessarily influence the linearly closest gene, their 3D organisation inside of the cell nucleus is crucial for the correct function of the genetic machinery.

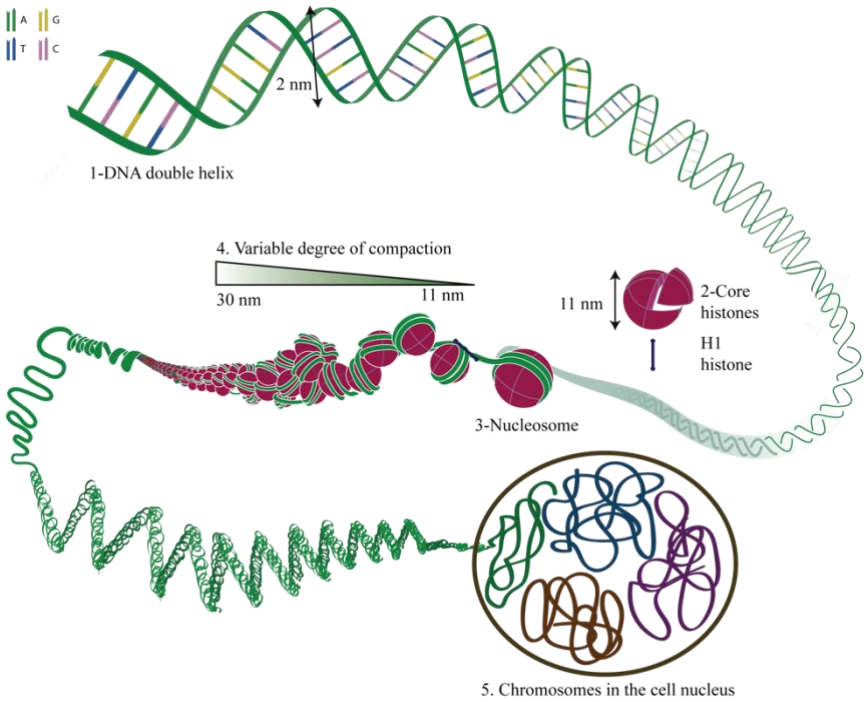


Figure adapted from Pierce, Benjamin. Genetics: A Conceptual Approach, 2nd ed.

Figure 1. DNA compaction in the nucleus from the double helix to chromatin. Schematic representation of the DNA structure and folding in the cell nucleus. (1) 2 nm wide DNA double helix structure, with the phosphate backbone in green and the four nucleotides in green (A), yellow (G), blue (C), and purple (T). (2) The four core histones present in the cell nucleus can aggregate together conforming a protein complex of a maximum with of 11 nm. Additionally, this complex can interact with a fifth histone, the Histone H1. (3) When the DNA helix is wrapped around a histone complex it conforms the nucleosome, a protein-DNA complex that increases the DNA compaction. (4) Different organisation of the nucleosomes can result in a variable degree of compaction of the chromatin, from the 11 nm width of a single nucleosome, to the 30 nm width organisation of many nucleosomes. (5) This arrangement is variable through the length of the chromosomes, and influences its disposition in the nuclear space. Figure adapted from (*Pierce 2012*).

Hence, DNA folding and unfolding in the cell nucleus must be highly efficient. Furthermore, the DNA needs to be accessible to the proteins associated with the replication, transcription and various regulatory processes of the cell, so it also needs to be extremely organised and dynamic. To this end, proteins interplay with the DNA at different genomic scale levels (*Wani, Boettiger et al. 2016*, *van Steensel and Furlong 2019*). At the nucleotide scale, DNA is associated with different sets of proteins to form the chromatin fibre. Specifically, segments of 145-147 bp of DNA wrap around

eight core histones (two copies of H2A, H2B, H3, and H4) (*Luger, Mader et al. 1997*) to form the basic unit of chromatin, the nucleosome (**Figure 1.2-3**). Nucleosomes are connected by free DNA strings of variable length named linker DNA, which are usually associated with linker histones. Linker histones are a group of histones which bind to nucleosomes by interacting with both their DNA and protein components. They modify the DNA exit angle from the nucleosome and help to neutralise the charge of the linker DNA, thereby affecting the level of compaction and accessibility of the chromatin fibre (*Klemm, Shipony et al. 2019*).

2 Structural organisation of the chromatin

Chromatin is arranged in a variety of conformations that ensure its proper compaction levels. The degree of compaction is dependent on the density of nucleosomes, which will be low and more dynamic on accessible chromatin, and high and stable in closed chromatin (*Schones, Cui et al. 2008*, *Deal, Henikoff et al. 2010*, *Ricci, Manzo et al. 2015*). Although still under debate, the chromatin appears to be organised in irregular nucleosomal organisation patterns, resulting in ranges of compaction of the chromatin that might vary between 5 and 30 nm, depending on the used experimental measure (*Finch and Klug 1976*, *Ou, Phan et al. 2017*, *Hsieh, Cattoglio et al. 2020*) (**Figure 1.4**). This variability would be suited for the sharp opening or closing of target genomic regions by modifying their compaction and thus, accessibility level. Interestingly, the percentage of accessible chromatin can be as low as the 2-3% of the whole genome in a given cell (*Thurman, Rynes et al. 2012*), covering ranges of non-continuous genomic regions. In consequence, chromosomes will show irregular compaction patterns that might shape their positioning in the cell nucleus (**Figure 1.5**).

Nevertheless, the activity and compaction stage of a given genomic loci is cell type and stage-dependent (*Lieberman-Aiden, van Berkum et al. 2009*, *Ricci, Manzo et al. 2015*), thus requiring reversible mechanisms to guide the switch. Among them we have ATP-driven chromatin remodelling complexes (remodellers) and the changes driven by modifications in the N-terminal tails of the histones conforming the nucleosome (**Figure 2**).

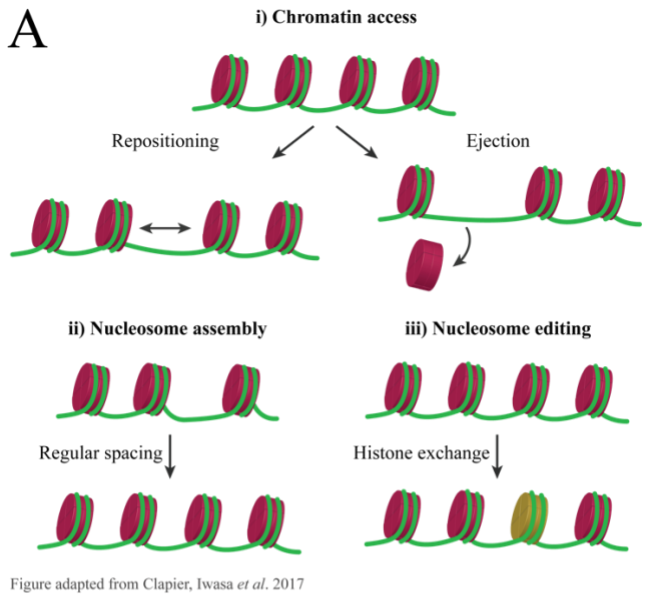


Figure 2. Mechanisms regulating chromatin compaction. (A) Functional classification of chromatin remodellers (remodellers). The classification englobes their role in: (i) Chromatin access, by which they modify the accessibility of the DNA; (ii) nucleosome assembly, that regulate the spacing of the nucleosomes; and (iii) nucleosome editing, that modulate the exchange of nucleosomes with histone variants. Figure adapted from (*Clapier, Iwasa et al. 2017*). (B) Representation of the

core histone complex. The four core histones (H3, H2B, H4, and H2A) are represented with different colours, and gather in pairs to conform the histone octamer. The histone tails of the upper half of the histone octamer are represented in grey, with two of the possible N-terminal modifications (methylation and acetylation).

Although it is not fully known how remodellers select their target nucleosomes, their function has been well characterised over the years (*Clapier, Iwasa et al. 2017*). Specifically, remodellers are mainly specialised in one of 3 functions: i) chromatin access, ii) nucleosome assembly and organisation, and iii) nucleosome editing (**Figure 2A**). Hence, they promote silencing or expression of genomic loci by the packing and unpacking of nucleosomal arrays (i and ii) and also by the turnover and exchange of canonical or variant histones (iii). By these means, remodellers affect nucleosome stability, factor recruitment, and exclusion, having an impact on the activity of the involved loci (*Clapier, Iwasa et al. 2017*).

Similarly, histone N-terminal tail modifications, like methylation and acetylation, add a second level of dynamicity to nucleosomes (**Figure 2B**). Histone methylation, for instance, affects to the binding affinity of numerous proteins, specifically by the individual or combinatorial methylation of Lysine 4 on histone H3 (H3K4), H3K9, and H3K27 (*Bartke, Vermeulen et al. 2010*). Depending on the direction in the affinity change and the proteins involved, some histone methylations are generally associated with active or inactive stages of the chromatin. For example, on the one hand, histone methylations like di/trimethylation of lysine 9 on histone H3 (H3K9me2/me3) and H3K27me3 are associated with different processes of chromatin condensation and hence with inactivity. On the other hand, H3K4me1 and H3K4me3, and H3K36me3 are associated with an active stage, specifically with the presence of enhancer elements, and transcription. Besides, histone acetylation marks are also associated with the activation of wrapped DNA locus by the loosening of its bound to the histone, specifically in loci containing regulatory elements. Altogether, these modifications and their combinations have profound effects on the activity sates of the chromatin, which result in changes in the chromatin organisation itself (*Siggens and Ekwall 2014*).

2.1 Organisation of the chromatin at the megabase scale

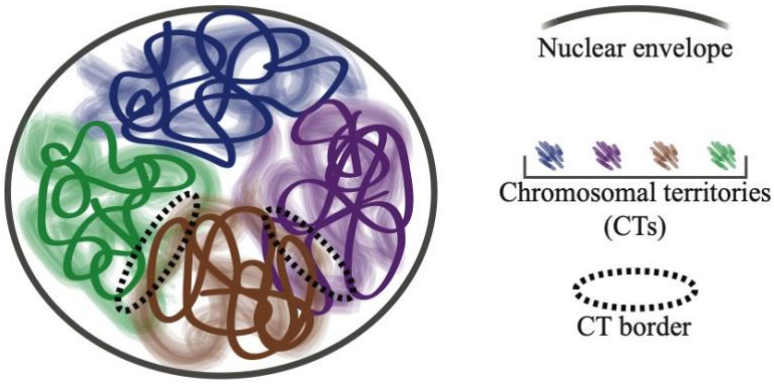


Figure 3. Chromosome positioning in the cell nucleus. Four examples of simplified chromosomes are represented in blue, purple, brown, and green. Their associated chromosomal territories are displayed in shadows of the same colour, and two of the peripheries of the brown chromosome appear highlighted by a dotted line circle.

In the cell nucleus, chromatin is hierarchised at different genomic scale levels. At the whole chromosome level, it is evident that chromosomes preferentially interact within themselves than between each other, leading to the formation of so-called chromosomal territories (CT) (*Cremer and Cremer 2010*). CTs are distinct nuclear volumes preferentially occupied by a single chromosome (**Figure 3**). Interestingly, within the nucleus, the CTs of long and gene-poor chromosomes tend to be located more towards the periphery of the nucleus. In contrast, the CTs of short and gene-richer chromosomes usually locate closer to the centre of the nucleus.

The borders between the CTs are characterised by a low density of chromatin, which enables the intermingling of loci from different chromosomes (*Ulianov, Gavrilov et al. 2015*), thus facilitating inter-chromosomal interactions. Some of these interactions have shown to be physiologically relevant and tend to form domains that are likewise co-occupied by specific chromosomal regions (*Maass, Barutcu et al. 2019*). Indeed, chromosomal arms can occupy distinct chromosomal territories (*Dietzel, Jauch et al. 1998*), which will likely compartmentalise in preferentially interacting regions. This

compartmentalisation can result in the observed polar distribution of gene-rich areas of the chromosomes towards the nuclear centre and of the gene depleted ones towards the periphery (*Kupper, Kolbl et al. 2007*).

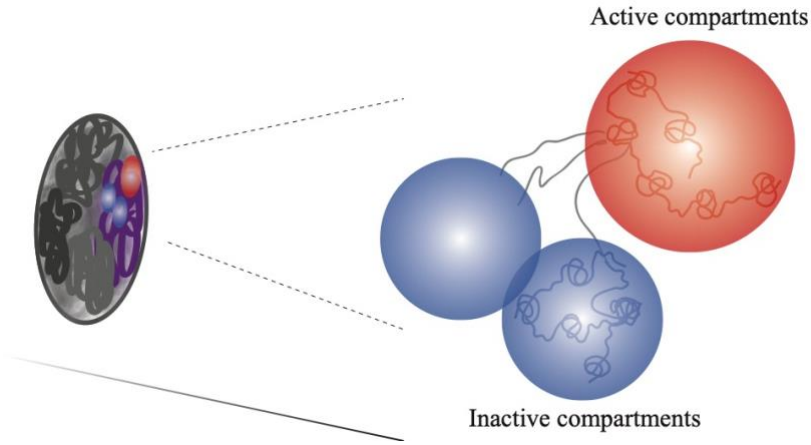


Figure 4. Compartments. As we move deeper into the organisation scale of the chromatin, the compartment arise. Compartments are megabase scale domains with specific activity, segregation, and epigenetic features. In this representation, active compartments are displayed in red, whereas inactive compartments appear represented in blue.

Given the resolution and the different types of experimental data used, compartments have been classified in groups that range from 2 to 6 according to different interaction, epigenetic and expression patterns (*Lieberman-Aiden, van Berkum et al. 2009*, *Filion, van Bemmel et al. 2010*, *Rao, Huntley et al. 2014*, *Vilarrasa-Blasi, Soler-Vila et al. 2019*). The classification usually involves different levels of active and inactive states. However, to date, most of the research groups divide them into two main categories that define mainly active (and open) and mainly inactive (and close) chromatin regions. Furthermore, these two categories overlap with the previously described euchromatin and heterochromatin (*Pueschel, Coraggio et al. 2016*). On the one hand, the heterochromatin is characterised by a higher density of nucleosomes and histone marks and variants associated with inactivity stages, resulting in a more compacted organisation. It tends to distribute more towards the periphery of the cell nucleus and associate with components of the nuclear envelope (*Schneider and Grosschedl 2007*). On the other

hand, the euchromatin is more accessible as a consequence of both the lower density of nucleosomes and the enrichment in acetylated histones (*Klemm, Shipony et al. 2019*) that characterise active chromatin regions. It tends to distribute more towards the centre of the cell nucleus, segregating from the heterochromatin. However, the genomic boundaries of these compartments are not consistent across different cell types, and cells can switch between 30% to 60% of their compartments as they get differentiated to other cell types and tissues (*Dixon, Jung et al. 2015*, *Schmitt, Hu et al. 2016*). Thus, the compartmentalisation of the chromosomes will be specific to the functional activity and stage of the cell (*Lieberman-Aiden, van Berkum et al. 2009*), segregating active and inactive areas, and in consequence, defining areas with ranges of active and absence transcription.

2.2 Organisation of the chromatin at the kilobase scale

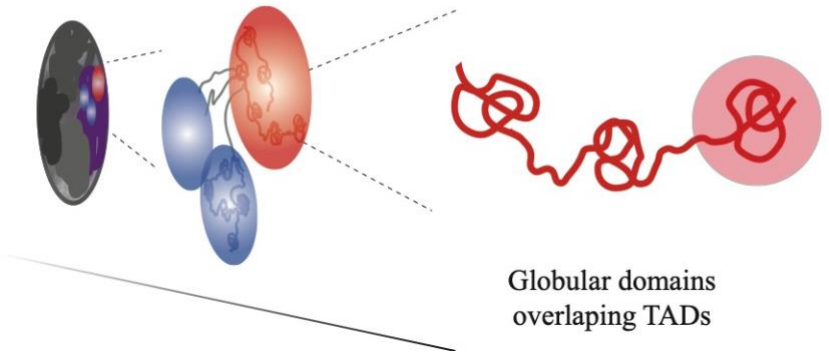


Figure 5. Topologically Associating Domains (TADs). As we move deeper into the organisation scale of the chromatin, TADs and globular domains appear. TADs are linearly contiguous fragments of the genome with a higher tendency to interact within each other than between themselves, and are usually represented as the globular domain highlighted by the red circle.

As we reduce the scale at which the genomic organisation is analysed, we find another layer of self-interacting domains named Topologically Associating Domains or TADs. TADs are linearly contiguous fragments of the genome with a higher tendency to interact within each other than between themselves (*Dixon, Selvaraj*

et al. 2012, *Nora, Lajoie et al. 2012*). Their borders are statistically detectable at the cell population level and not ubiquitously defined in individual cells from the same type (*Bintu, Mateo et al. 2018*). Hence, each cell from a population will most likely have globular structures whose start and end coordinates partially overlap the ones of a TAD (**Figure 5**). TADs can have a size that ranges from 40 kb to 3 Mb (*Rao, Huntley et al. 2014*) and their definition is influenced by the resolution of the experiment and the tool used to detect them. This explains the lack of consensus between experiments at the time to define the exact borders of the TADs. However, TAD boundaries, unlike compartments, are relatively conserved between species and tissues. Thus, suggesting that the interactions within the regions enclosed inside are functionally more relevant for the correct function of the cell than the ones involving other TADs (*Lieberman-Aiden, van Berkum et al. 2009*, *Ulianov, Khrameeva et al. 2016*).

Indeed, the genomic regions delimited inside the same TAD generally show similar trends of histone signatures, expression levels, and replication timing. Additionally, their boundaries overlap with those of replication domains (*Pope, Ryba et al. 2014*, *Bouwman and de Laat 2015*). They also facilitate cell-type-specific enhancer-promoter interactions (*Bonev, Mendelson Cohen et al. 2017*) and insulate them from unwanted interactions with elements from contiguous TADs (*Lupianez, Kraft et al. 2015*). However, experiments disrupting TAD patterns show different outcomes. Sometimes cells can survive without most of the TADs (*Nora, Goloborodko et al. 2017*) or are minorly affected after rearranging some of them (*Ghavi-Helm, Jankowski et al. 2019*), other times they lose strength in their response to external stimuli (*Stik, Vidal et al. 2020*), and in others the disruption of specific TADs is enough to drive to malfunction (*Lupianez, Kraft et al. 2015*). Thus, suggesting that different factors might be involved in the formation of the TAD patterns and that each of them might have a variable degree of importance for the function and sensitivity of the cellular processes.

In mammals, TAD borders are characterised by the presence of the structural maintenance of chromosome (SMC) complex and the CCCTC-binding factor (CTCF) (*Szabo, Bantignies et al. 2019*) (**Figure 6**). SMC complexes are ring-shaped proteins involved in the formation and further enlargement of chromatin loops, among

other functions (*Sedeno Cacciatore and Rowland 2019*). CTCF on the other hand, is a zinc finger protein that binds to the genome to exert as a transcriptional activator, repressor, or insulator (*Kim, Yu et al. 2015*). Specifically, it binds to non-palindromic consensus sequences that are usually found at both of the TAD borders in a convergent orientation (*de Wit, Vos et al. 2015*).

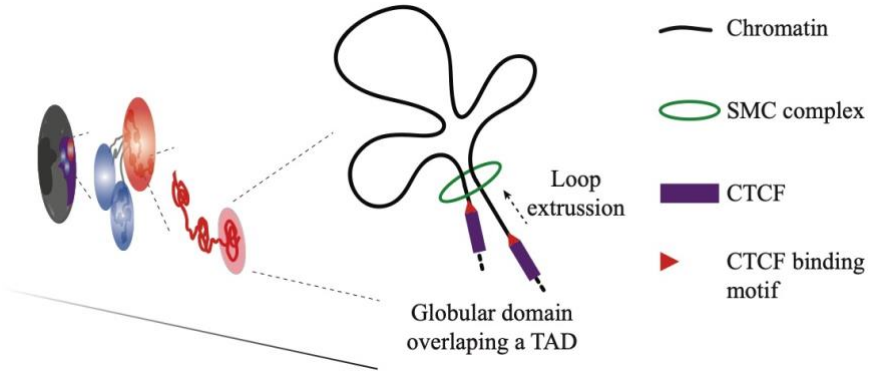


Figure 6. TADs in detail. As we move deeper into the organisation scale of the chromatin, the features characterising globular domains appear with more detail. These features are proposed to be the result of a loop extrusion process that is driven by the SMC complex. It finishes when the SMC complex collides with two CTCF proteins oriented towards the extruding loop. Thus defining the boundary of the globular domain.

There is increasing evidence suggesting that both SMC complex proteins and CTCF are involved in TAD patterns formation. Concretely, a mechanism called loop extrusion (*Alipour and Marko 2012*) is rapidly earning support in the scientific community. According to it, an SMC complex ring would load to the DNA by the mediation of the heterodimeric complex NIPBL-MAU2, then surround a small chromatin loop, and push it out from both edges (or extrude it) until colliding with a convergently oriented CTCF protein (*Davidson, Bauer et al. 2019*). Concordantly TAD patterns are generally lost in populations of cells depleted of CTCF or cohesin (a protein member of the SMC complex). However, CTCF depletion is not sufficient to remove all TADs (*Nora, Goloborodko et al. 2017*, *Rao, Huang et al. 2017*). This suggests that processes related with transcription and promoter-enhancer looping might be involved in the formation of the remaining ones.

2.3 Organisation of the chromatin at the bp scale

Chromatin loops are chromatin structures by which two linearly distant regions (defined as anchors) converge in the 3D space (**Figure 7**). Loops tend to bring together regulatory and target elements (*Greenwald, Li et al. 2019*) separated by genomic distances that range from 40 kb to 3 Mb (median of 185 kb). Previous studies comparing human and mouse cell lines have shown that loops can be conserved between cell lines (55-75%) and species (50%) (*Rao, Huntley et al. 2014*).

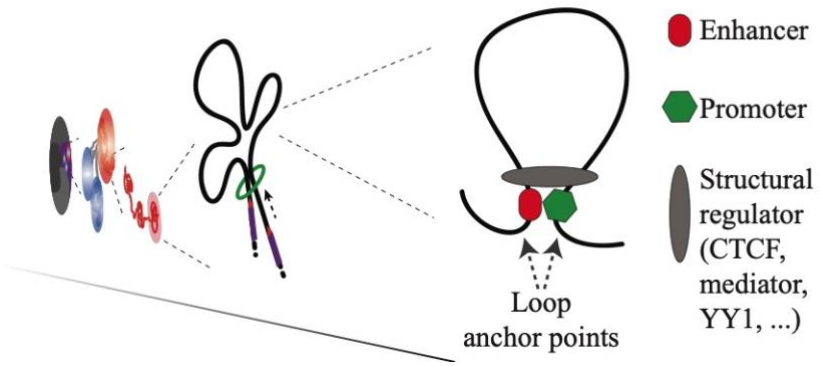


Figure 7. Loops. As we move deeper into the organisation scale of the chromatin, loops emerge. Loops, as their name state, are chromatin folding that result from the bending of the chromatin to bring two distant loci, usually containing regulatory elements like enhancers and promoters, together. This process is usually regulated by a structural regulator protein like CTCF, mediator, and Yin Yan 1 (YY1), among others.

Most of the loop anchors present CTCF and the cohesin subunits SMC3 and RAD21 (86%, 87%, and 86% of them respectively) (*Rao, Huntley et al. 2014*). However, although loops enclosed by CTCF sites can be involved in coordinating the expression of the contained genes, CTCF-binding sites are generally far from promoters in the human genome (*Kim, Abdullaev et al. 2007*). This disposition suggests that other structural proteins might be involved in their direct looping. Indeed, Yin Yan 1 (YY1), a zinc finger protein like CTCF, is involved in looping processes that promote enhancer-promoter interactions (*Weintraub, Li et al. 2017*). Mediator, a multi-subunit protein complex, has also been found in loop anchors connecting enhancers and promoters of actively

transcribed genes in a cell-type-specific manner. Mediator facilitates transcription by bounding the transcription factors attached to the enhancer sequences with the transcription machinery assembled at promoters (*Kagey, Newman et al. 2010*, *Soutourina 2018*).

The correct control on the formation and maintenance of these structures is mandatory for the functioning of the cell and aberrant loop formation can drive to cancer and diseased processes that are related with abnormal enhancer-promoter interactions (*Norton and Phillips-Cremins 2017*).

2.4 Promoters, enhancers, and super-enhancers

Promoters and enhancers are critical players in the regulation of the specific subsets of genes needed during cell function, interplay and survival. Historically, promoters have been classified both by their positioning at 50 bp around the transcription start site of the downstream gene and by their role recruiting the RNA Polymerase II (*Andersson and Sandelin 2020*). They also define the preferential direction of transcription and are usually characterised by H3K4me3 marks. On the other hand, enhancers are classified as elements located farther away from their target gene or genes (even up to 1Mb) and by their role in modulating target gene expression. They usually are characterised by a high H3K4me1 to H3K4me3 ratio and activity-dependent presence of H3K27ac (*Andersson and Sandelin 2020*). Nevertheless, both promoters and enhancers can have different ranges of mixed enhancer and promoter role, and similarly occupy nucleosome depleted regions that facilitate the access of proteins, like chromatin regulators, to the DNA (*Lai and Pugh 2017*).

Enhancers have also been grouped into additional categories according to their arrangement with other enhancers. The term super-enhancer for example, generally defines groups of enhancers proximally located in the linear genome and highly enriched in the occupancy of transcriptional coactivators like Mediator. This definition does not involve any functional property for super-enhancers, and the different classifications criteria associated with the group makes their definition a bit loose (*Pott and Lieb 2015*).

Other classifications however, have tried to address this issue by adding a functionality level. For example, the term enhancer hub includes a spatial dimension and defines enhancer enriched 3D domains that have shown to work as functional units (*Miguel-Escalada, Bonas-Guarch et al. 2019*).

2.5 Wrap up

Altogether, it is clear that to maintain proper cellular function and avoid diseased stages, enhancer-promoter interactions need to be finely regulated. Structural regulators play a crucial role in this processes by facilitating their looping interactions and further aggregation of sets of these loops into the frequently interacting areas defined as TADs (*Clapier, Iwasa et al. 2017*, *Soutourina 2018*, *Szabo, Bantignies et al. 2019*). TADs can, in the same way, arrange together in linearly distant groups by random movement in the cell nucleus, homotypic attraction, or activity stage, for instance (*Fraser, Ferrai et al. 2015*, *Robson, Ringel et al. 2019*). Some of them can gather in the same compartment and similarly, active compartments show a tendency to colocalise. This intermingling is present at all the levels of the chromatin organisation and errors altering this arrangement can lead to diseased stages of the cell (*Maass, Barutcu et al. 2019*). Hence, tools for the study of the chromatin organisation hold the potential to characterise the mechanisms behind specific gene expression and regulation, and so the ones that lead to disease and malfunction.

3 Experimental procedures for the analysis of chromatin organisation

The analysis of the organisation of the chromatin is mainly divided into two approaches: imaging-based methods and molecular biology-based methods.

3.1 Imaging methods

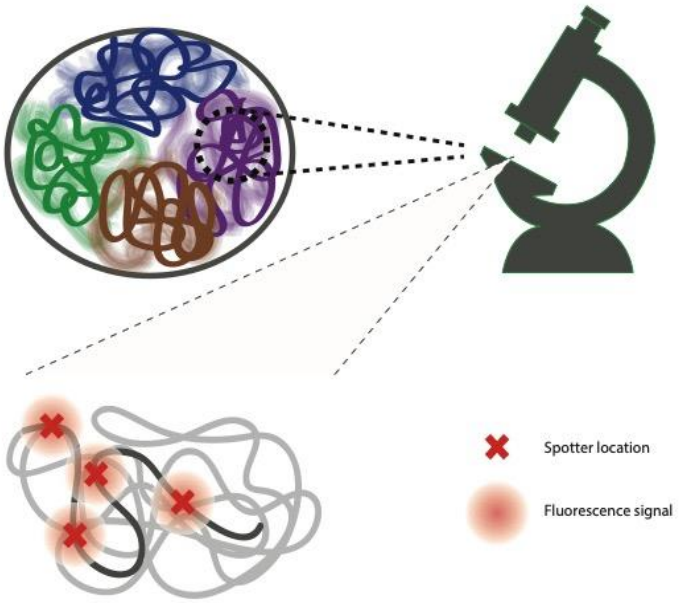


Figure 8. Simplification of the imaging-based methods approach. Overall, this methods involve the observation of the chromatin organisation by different types of microscopes. To track the location of specific genomic loci, cells are previously treated to introduce an spotter that binds to these specific loci to facilitate a fluorescent signal. Spotters can be nucleic sequences like probes or oligos, or protein complexes like CRISPR-cas9 for instance.

Imaging methods rely on the usage of different types of microscope modalities for the analysis of the chromatin organisation and, as such, they are limited by the resolution of the microscope. Nevertheless, the achievable resolution ranges from 250 nm with light microscopes to 10-30 nm with super-resolution microscopy. The majority of the microscope methods designed to investigate chromatin are based on fluorescence *in situ* hybridisation (FISH) protocols (*Bauman, Wiegant et al. 1980*) that rely on the usage of labelling probes to track the location of genomic loci and require the fixation of the cells prior to the visualisation step. Nevertheless, there are alternatives to FISH methods that do not require the DNA labelling and fixation steps, like CARGO for example (*Gu, Swigut et al. 2018*).

3.1.1 FISH-based methods

FISH-based methods are widely used for the analysis of the genomic 3D organisation. **FISH** is one of the first imaging approaches introduced for the 3D location of specific loci. FISH uses specifically designed fluorescent DNA probes of few kilobases that bind to the genomic DNA by complementarity, pointing the nuclear locations in which they attach. The experimental set up was initially limited to the introduction of small sets of probes that were suited for the detection of long-range interactions (*Bauman, Wiegant et al. 1980*, *Bienko, Crosetto et al. 2013*). Nevertheless, new advances in microscopy and molecular biology methods have led to the development of a wide variety of methods aimed for resolving different needs (*Volpi and Bridger 2008*). **Cryo-FISH** for instance, has aimed for a greater accuracy by a step of cryo-sectioning the cells (in layers of ~100-200 nm). This helps to set finer boundaries for the location of the signal (*Xie, Lavitas et al. 2010*). Other methods have aimed for the optimisation of the process. **HIPMap** for example, redesigned the protocol in a high throughput way. For that, it tags thousands of cells in a 94- or 384-well plate and images all them at once. It then uses a specific bioinformatic tool to place all the cell nuclei and track the relative positions of the genes within themselves and the nuclear border (*Shachar, Pegoraro et al. 2015*).

One of these adaptations, **Oligopaint** (*Beliveau, Joyce et al. 2012*), has done a step forward laying the foundations for a new set of techniques. Oligopaint is capable of locating more loci in the genome than most of the previous FISH approaches and to even trace different ranges of continuous genomic coordinates. It involves the bioinformatic design and production of thousands of oligos that can reach a density of around 10 per kb. This design in conjunction with super-resolution microscopy (**OligoSTORM**) (*Beliveau, Boettiger et al. 2017*) improved the quality of the data obtained from 25 to 100 nm resolutions, but at the same time allowed the analysis of regions at the megabase scale.

Further optimisations adapted the Oligopaints method by sequential rounds of labelling and diffraction limited imaging (*Wang, Su et al. 2016*). In this way, a set of primary oligos are hybridised to the

genomic sequences of interest in order to track them in space. Then, a series of steps of photobleaching and hybridisations are used to sequentially label and image the loci of interest by the use of secondary probes. Specifically, each primary oligo is labelled by appending a specific barcode sequence that facilitates the selective hybridisation of the secondary oligos. With this method, they were able to track the position of tens of TADs from different chromosomes. Additionally, the sequential Oligopaints method has also been used in conjunction with super-resolution (*Bintu, Mateo et al. 2018*, *Nir, Farabella et al. 2018*) allowing to sequentially label continuous genomic coordinates of the genome with fixed size steps of 30 kb (*Bintu, Mateo et al. 2018*) or variable size steps (10 kb-1.8 Mb) (*Nir, Farabella et al. 2018*). Thus, unveiling the folding of contiguous regions of the genome at the level of single gene, loops, TADs and compartments. Another technique that adapted sequential imaging is **ORCA** that allowed to increase its resolution to 2-10 kb and tagging regions from 100 to 700 kb long (*Mateo, Murphy et al. 2019*).

Other flavours of Oligopaints-based microscopy have aimed at the joint detection of the positioning and transcriptional activity of loci. In this mater, **Hi-M** simultaneously tracks the 3D position of the tagged loci and their transcriptional activity, covering regions of 400kb at an average resolution of 17 kb (*Cardozo Gizzi, Cattoni et al. 2019*). Other approaches have aimed for the automation of the process. Specifically, **OligoFISSEQ** (*Nguyen, Chattoraj et al. 2020*) has taken a massive step towards the high throughput imaging and tracing of genomic loci in thousands of cells. It uses a combination of fluorescence in situ sequencing (FISSEQ) (*Lee, Daugharthy et al. 2015*), a method for in-situ RNA sequencing, and Oligopaints. OligoFISSEQ can be used for characterising multiple genomic loci at high resolution or chromosomes at a lower resolution. Most recently, **DNA-MERFISH (***Su, Zheng et al. 2020***)** has also allowed the genome-wide tracking of 1,000 genomic loci together with their associated transcripts**.** Altogether, FISH-based techniques are rapidly evolving towards a higher resolution and throughput, making it even possible to jointly observe chromatin 3D organisation and expression. However, these methods are limited by the cell fixation step, which hampers the analysis of the dynamic behaviour of the chromatin.

3.1.2 Alternative methods

As an alternative to FISH-based methods, some approaches are focused on the usage of DNA-binding proteins that do not require previous fixation of the cells. Among them, CRISPR-cas9 system-based methods are gaining in popularity (*Lakadamyali and Cosma 2020*). These methods rely on the prior integration in the genome of an endonuclease-deactivated Cas9 protein (dCas9) tagged with an enhanced green fluorescent protein. Then the sgRNA that guides the attachment of the dCas9 system is introduced in the cell. Since the introduction of the specific gRNAs is the limiting factor in these experiments, most of them have focused in the detection of regions containing sequences of repetitive elements (*Chen, Gilbert et al. 2013*). However, further adaptations like CARGO (*Gu, Swigut et al. 2018*) have allowed locating non-repetitive genomic sequences of 5 kb and 2 kb length by improving the delivery of the gRNAs. In this way, Gu and colleagues were able to study the dynamic behaviour of cis-regulatory elements.

3.2 Molecular biology methods

Molecular biology methods focus on the retrieval and further sequencing of interacting chromatin regions located in the proximity of other loci or proteins in the cell nucleus. Most of the molecular biology methods have been adapted from chromatin conformation capture (3C) technique (*Dekker, Rippe et al. 2002*). 3C-based approaches are based in the chromatin cross-linking, digestion (either with Restriction Enzymes, sonication of both methods), and proximity re-ligation of the loose ends to produce chimeric DNA molecules (**Figure 9**). However, some variations avoid this step by different labelling strategies. In both methods, the chimeric or barcoded DNA molecules are amplified by PCR, sequenced, and mapped to different genomic coordinates relating the regions that colocalised in the genome. Then, the genome is fragmented into specific length bins depending on both the sequencing depth and the length of the digested fragments. Finally, the number of times in which two bins coincide in the same chimeric read are counted and assigned as frequencies of interactions between the involved genomic loci (*Kempfer and*

Pombo 2020). We can further classify these methods by the type of information they provide.

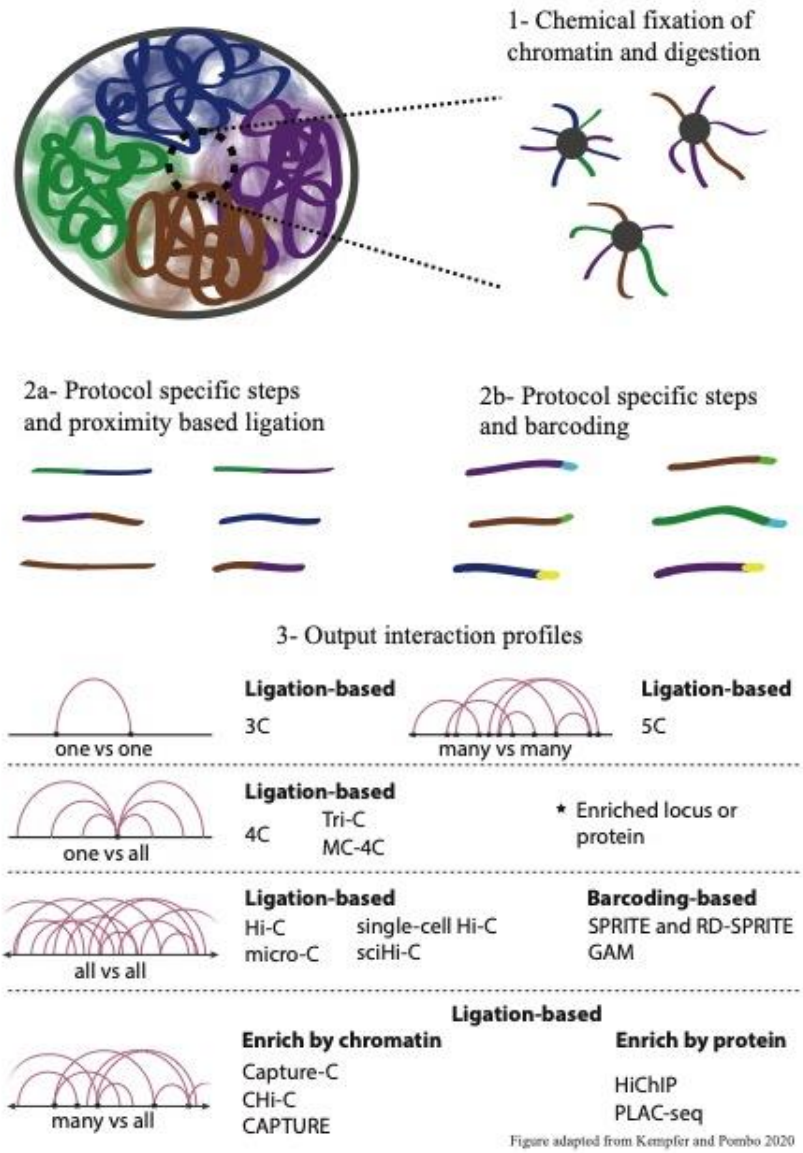


Figure 9. Simplification of the molecular biology methods. 1) Most of molecular biology methods involve the chemical fixation of the chromatin followed by digestion with restriction enzymes or sonication. 2) After a set of technique-specific steps, fragments can be related, either by proximity ligation or barcoding, as colocalising (or interacting) between each other. 3) These technologies provide interaction profiles that involve different subsets of genomic regions. Figure adapted from (*Kempfer and Pombo 2020*).

3.2.1 One vs one

Some experiments focus on the interactions involving an already known locus (viewpoint) or defined genomic region. **3C** for example, retrieves interactions involving two previously targeted loci (*Dekker, Rippe et al. 2002*), returning the number of times both loci are found together in a population of cells. It produces a *one vs one* interaction profile whose importance can be weighted by comparison with a control locus.

3.2.2 Many vs many

Chromosome conformation capture carbon copy (**5C**) (*Dostie, Richmond et al. 2006*) is a tool suited to obtain high resolution interacting matrices in a set of continuous loci of interest. 5C focuses on the retrieval of interactions within a contiguous genomic region of interest, by capturing all or most of the chromatin fragments located inside. Thus, it is classified as a *many vs many* approach. Its resolution is dependent on the size of the produced restriction fragments inside of the focus regions, and on experimental limitations that come with the requirement of capturing all of them.

3.2.3 One vs all

The scope of *one vs one* experiments can be increased by looking at all the interactions involving a viewpoint. In this way, circular chromosome conformation capture (**4C**) provides a *one vs all* picture, in which we can track how frequently a locus of interest interacts with the rest of the genome at resolutions of few kilobases (*Simonis, Klous et al. 2006*, *van de Werken, de Vree et al. 2012*). The 4C approach has been enhanced to retrieve interactions between groups of 3 or more genomic loci. **Tri-C** (*Oudelaar, Davies et al. 2018*) for example, has reduced the size of the digested fragments to fit more of them in the strand length optimal for Illumina sequencing. On the other hand, multi-contact 4C (**MC-4C**) (*Allahyar, Vermeulen et al. 2018*, *Vermeulen, Allahyar et al. 2020*) has taken advantage of long-read sequencing to increase the length of the sequenced chimeric reads in a way that can retrieve interactions between groups of 3 or more genomic loci. Both

technologies reach resolutions of few kilobases and can be treated to obtain matrices of pairwise interactions enriched around the viewpoint.

3.2.4 All vs all

Some experiments, classified as *all vs all* approaches, have been designed for the detection of unbiased whole-genome interactions. From these methods, **Hi-C** (*Lieberman-Aiden, van Berkum et al. 2009*, *Rao, Huntley et al. 2014*) is the most popular one. The Hi-C protocol relies on the labelling with biotin of the loose ends of the digested DNA fragments. This labelling allows the later retrieval and enrichment of the fragments that were re-ligated (and thus still hold the biotin label) by using streptavidin beads. Some of the adaptations of Hi-C, like **single-cell Hi-C** (*Nagano, Lubling et al. 2015*) and single-cell combinatorial indexed Hi-C (**sciHi-C**) (*Ramani, Deng et al. 2020*), have allowed moving from the population-based interaction data to the single-cell interaction data. Hence, addressing the subset of interactions that are present at once in a single cell, either by isolation of single cells or tagging with unique barcodes, respectively.

Other methods, though, have focused on the retrieval of interactions at a higher resolution, like **micro-C** (*Hsieh, Weiner et al. 2015*, *Hsieh, Fudenberg et al. 2016*). Micro-C uses a micrococcal nuclease, instead of sonication or restriction enzymes, to achieve single nucleosome resolution (~200 bp). Methods like split-pool recognition of interactions by tag extension (**SPRITE**) (*Quinodoz, Ollikainen et al. 2018*), have removed the proximity ligation of the digested loose ends, and instead rely in series of steps of dilution, tagging and, mixing. This process promotes the equal barcoding of the chromatin complexes that are maintained together through all the process, giving information of multiple interacting regions at few kilobases resolution. Further adaptations have allowed setting this method also to track interactions involving RNA (**RD-SPRITE**) (*Quinodoz, Bhat et al. 2020*) and single cells (**scSPRITE**) (*Arrastia, Jachowicz et al. 2020*).

Lastly, genome architecture mapping (**GAM**) (*Beagrie, Scialdone et al. 2017*, *Beagrie, Thieme et al. 2020*) also avoids the re-ligation

step at the time to get the co-localisation between genomic areas. Instead, GAM sections cells in 220 nm thin layers to separate more easily *in* space the interactions found on each of them. Then it relies on a mathematical model named SLICE to assess the degree of co-localisation of the different genomic regions at tens of kilobases resolution.

3.2.5 Many vs all

The last type of 3C-based experiments retrieve interactions involving dispersed loci with the rest of the genome, returning profiles defined as *many vs all*. **Capture-C** (*Hughes, Roberts et al. 2014*), Capture Hi-C (**CHi-C**) (*Mifsud, Tavares-Cadete et al. 2015*), and Promoter Capture Hi-C (PCHi-C) (*Schoenfelder, Javierre et al. 2018*), for example, use biotinylated RNA probes to pull-down interactions involving a set of viewpoints of interest, which can be as many as of thousands of them. These experiments are enriched for the interactions involving the loci of interest. In this way, they reduce costs in the sequencing of interactions that are not initially in the scope of the designed experiment and reach resolutions of few kilobases.

On the other hand, methods like **HiChIP** (*Mumbach, Rubin et al. 2016*) and Proximity Ligation-Assisted ChIP-seq (**PLAC-seq**) (*Fang, Yu et al. 2016*) are designed for the retrieval of interactions involving a protein of interest. These methods use ChIP to retrieve the re-ligated chromatin complexes where the protein of interest is present. Interestingly, CRISPR based technologies are also finding their way in molecular biology methods for the analysis of the chromatin. For example, CRISPR affinity purification in situ of regulatory elements (**CAPTURE**) (*Liu, Zhang et al. 2017*) retrieves interactions involving a set of locus of interest. Concretely, it uses a biotinylated engineered dCas9 which allows the recovery of protein, RNA, and DNA complexes associated with the target locus. This process relies on previously designed sgRNAs that will load into the dCas9 and target it to a specific locus or repetitive DNA sequences.

As can be seen, molecular biology methods are quite diverse, what is not surprising given that they do not require any specific machinery or installations apart from the ones usually found in a

standard laboratory. This situation promotes the usage and evolution of the technology together with the improvement of the data retrieval, which lately has shown to significantly agree with observations measured by imaging (*Bintu, Mateo et al. 2018*, *Nir, Farabella et al. 2018*, *Cardozo Gizzi, Cattoni et al. 2019*).

4 Analysis of molecular-biology-based chromatin organisation experiments

As all experimental techniques, molecular biology methods for the detection of the chromatin structure are not exempt from experimental biases. Furthermore, the fast evolution of the field is resulting in the development of more ambitious approaches, that sometimes need specialised tools for the contextualisation, filtering and normalisation of the data. For this reason, many of the new molecular biology methods for chromatin architecture come accompanied by a well-defined set of instructions or a specific bioinformatic tool for the treatment and normalisation of the experimental results. With these tools, they try to remove most of the biases of the experiments in an effort to make the obtained information as reliable as possible. SPRITE, 4C, and Tri-C are examples of techniques that provided an innovative way to normalise their data at the time of the release. The authors of GAM and MC-4C, on the other hand, went a step further and designed a specific bioinformatic tool for the treatment and further analysis of their datasets.

Meanwhile, some tools have become popular enough to see the release of different normalisation and treatment approaches developed by different groups. Hi-C and its adaptations, for instance, have a long list of alternatives regarding the filtering and normalisation of the data. Each of them has its own approach for the normalisation step and try to remove biases specific for different data sources. To name some, ICE (*Imakaev, Fudenberg et al. 2012*), HiCNorm (*Hu, Deng et al. 2012*), HiC-Pro (*Servant, Varoquaux et al. 2015*), HiCUP (*Wingett, Ewels et al. 2015*), Juicer (*Durand, Shamim et al. 2016*), and OneD (*Vidal, le Dily et al. 2018*) are some of the most common tools and approaches. Promoter Capture protocols have also seen the rise of some tools for the treatment and

detection of significant interactions in the datasets. Examples of these tools are CHiCAGO (*Cairns, Freire-Pritchett et al. 2016*) and ChiCMaxima (*Ben Zouari, Molitor et al. 2019*). Other tools have focused in the differential analysis between datasets, like diffHic (*Lun and Smyth 2015*), FIND (*Djekidel, Chen et al. 2018*), and Selfish (*Ardakany, Ay et al. 2019*) for HiC, and Chicdiff (*Cairns, Orchard et al. 2019*) for capture Hi-C.

5 Bioinformatic methods for the 3D representation and analysis of chromatin structure

The major issue of working with chromatin interaction data is the loss of the 3D perspective. This loss makes it difficult to contextualise the interaction profiles in the original 3D organisation of the chromatin. In this mater, tools for the chromatin 3D modelling emerge as an essential instrument, for both the contextualisation and further analysis of the genome architecture. Chromatin 3D models have a double function. The first one is to help in the 3D visualisation behind the interaction profiles, which can help researchers to better interpret their data. The second one is the inclusion of the data in the third dimension, which allows analysing the genome in the context of its distribution in space. Most modelling approaches subdivide the genome in chunks by some underlying features or a defined genomic length. Then each chunk is represented by connected points or spheres, although some methods go a step further and model them as elements composing a polymer (*Oluwadare, Highsmith et al. 2019*). These particles are then constrained by a series of parameters or physics rules that will define how they fold and interact with the rest of the particles of the model and the simulated environment.

Chromatin 3D models can be divided into two categories by the input data used. *Ab initio* models use as input statistical physics and features that shape the behaviour of the simulated chromatin. On the other hand, data-driven models are focused on the treatment and transformation of experimental data to reliably reconstruct its 3D organisation. In this way, *Ab initio* models will aim at the understanding of the processes shaping genome folding, while data-driven models will have as focus the more refined analysis of the

represented chromatin (*Marti-Renom and Mirny 2011*, *Lin, Bonora et al. 2019*, *Bendandi, Dante et al. 2020*)

5.1 *Ab initio* modelling

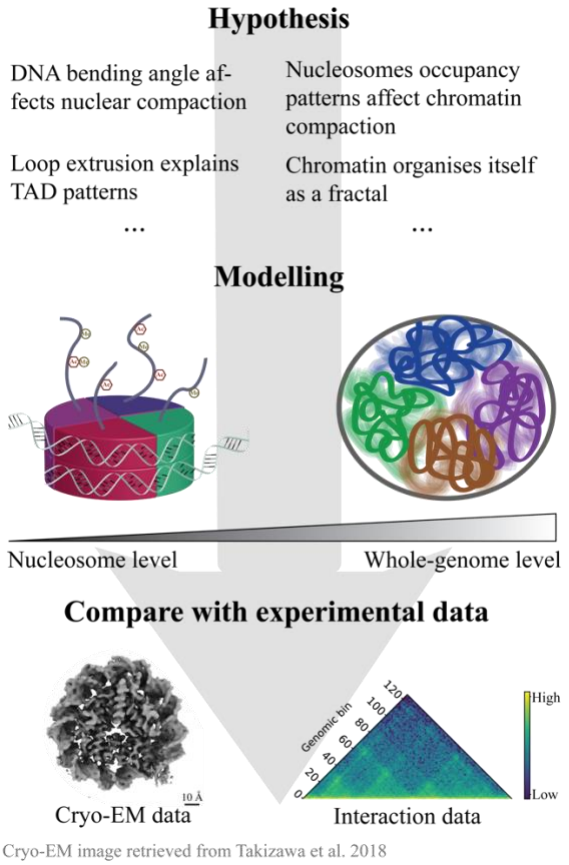


Figure 10. *Ab initio* modelling workflow. The forces driving chromatin organization are inquired by testing the hypotheses or accepted truths involving specific and well-defined physical properties of the modelled environment. These hypotheses might focus on different genomic scales, from the nucleosome level to the loops, TADs or the whole genome level. Once the model is obtained, experimental data is used to test how well do the imposed rules explain the ground truth.

Ab initio chromatin models (**Figure 10**) aim to reproduce and understand specific features of the chromatin by applying a conjunction of already known, and sometimes also hypothesised, properties of the elements that conform the chromatin. These methods have a broad scope and can aim from the analysis of the

organisation of the DNA at the nucleosome level to the study of the chromatin organisation at the level of TADs, chromosomes, and the whole genome (*Bendandi, Dante et al. 2020*). Methods that are more specific to the analysis of the DNA at the nucleosome level focus in the understanding of forces related with the bending angle of the nucleosomes (*Koslover, Fuller et al. 2010*), the presence of the linker histone (*Pachov, Gabdoulline et al. 2011*), and the molecular mechanic force fields associated to nucleotides (*Cheatham and Case 2013*) among others. The bigger it is the scope of these models, the more elements it will have to take into account for the modelling step, until reaching a scale that is not computationally feasible with such a level of the detail. For this reason, models aiming to analyse large chromatin regions often simplify the structure and biological factors involved at the small scales, like single atomic or nucleosome dispositions, that are not the focus of those experiments.

Models aiming to analyse large chromosomic regions use properties inferred from experimental observations to represent the chromatin fibre. Usually, the chromatin fibre is modelled assuming a 30 nm packing conformation (*Finch and Klug 1976*) and the behaviour of the bead-spring polymer models (*Kremer and Grest 1990*, *Rosa and Everaers 2008*). Polymers, as chromatin, are large macromolecules composed of chains of repetitive subunits. This organisation together with the added mass of the macromolecule results in specific, and extensively studied, physical properties regarding their toughness, elasticity, and behaviour. Hence, the genome can be treated as a polymer in which each subunit represents a genomic chunk with well-defined start and end coordinates (*Lin, Bonora et al. 2019*). At the chromatin organisation scale, polymer folding can be modelled as a fractal globule (*Grosberg, Nechaev et al. 1988*) and as an equilibrium globule (*Mirny 2011*) for instance. Indeed, the fractal globule organisation was proposed based on the first genome-wide chromosome interaction maps (*Lieberman-Aiden, van Berkum et al. 2009*). This organisation is the consequence of a condensation event in which regions of the polymer chain are prevented from passing across each other, resulting in a configuration that can be rapidly unfolded and refolded. This configuration change would promote the accessibility of specific regions of the polymer as it happens in the nuclear chromatin during

transcriptional activation events that promote the accessibility of the DNA.

Methods following these approaches have helped to proof that loop extrusion processes could be sufficient to drive chromatin compaction (*Goloborodko, Marko et al. 2016*) and form chromosomal domains (*Fudenberg, Imakaev et al. 2016*), and that coregulated genes could colocalise in the nuclear space shaping chromatin organisation (*Di Stefano, Rosa et al. 2013*). Similarly, these methods can also be useful in the analysis of specific loci. In fact, a genome folding strategy based on the bridging between inferred binding sites has already been used to recover most of the structural organisation of specific loci (*Chiariello, Annunziatella et al. 2016*, *Chiariello, Bianco et al. 2020*).

5.2 Data-driven modelling

Data-driven modelling (**Figure 11**) aims to represent the chromatin organisation in a way in which its 3D folding can be contextualised, allowing further analysis based on the spatial conformation of the modelled regions and their relative distances. To do so, these methods usually represent the genome as a concatenation of points or spheres that subdivide the genome in chunks. The resolution of the experiment and the computational workload are the limiting factors at the time to define the length of the chromatin fibre represented in these chunks. For this reason, models aiming to analyse the chromatin organisation of long genomic coordinates, like the whole human genome, usually have low resolutions of about a megabase. On the other hand, models focused on the sharp analysis of previously selected regions of interest typically have resolutions of around few kilobases, that are closer to the limit defined by the experiment (*Serra, Di Stefano et al. 2015*).

Once the representation is set, most data-driven models need a way to score the input interaction data. In this way, they can infer how well do the output 3D distances between the model particles represent the input interaction data. These scoring functions are mainly designed by taking into account restraints inferred from the: interaction data, additional experimental observations, and physic properties of the chromatin (*Serra, Di Stefano et al. 2015*). The

restraints inferred from interaction data usually follow an inverse relationship with the interaction frequency and are modified by an exponent parameter. Additional experimental restraints can include information about the nuclear dimensions and the distribution of specific chromatin structures on it, among others. Lastly, physics-based restraints usually take into account the connectivity of the polymer, the overlapping degree between particles, and the bending rigidity between consecutive particles (*Serra, Di Stefano et al. 2015*).

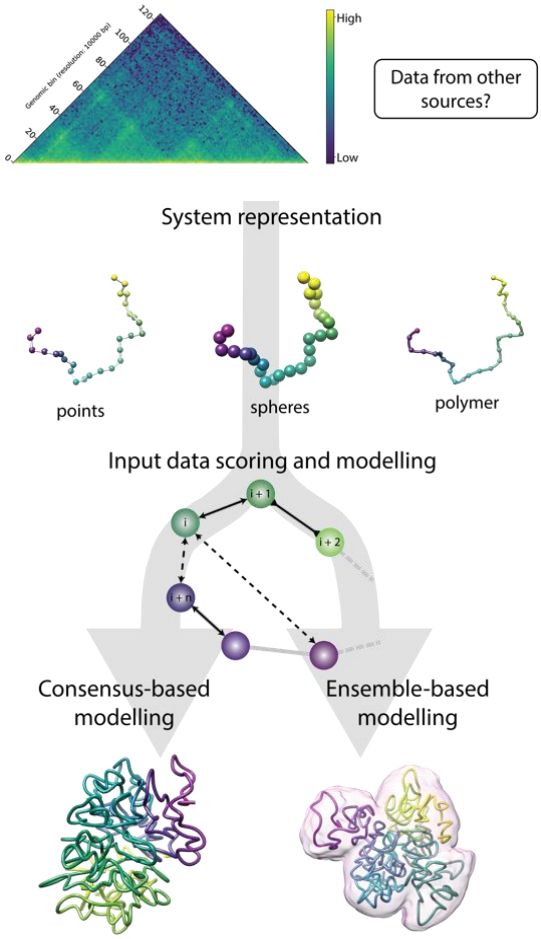


Figure 11. Data driven modelling workflow. 3C-based interaction data, and sometimes other sources of chromatin organisation data, are used to restrain in space the modelled system. These systems are usually represented as a chain of points or spheres, or as a polymer. Then, the inferred restraints are applied and scored in order to select the model (consensus modelling) or ensemble of models (ensemble-based modelling) that best represent the input data.

Finally, the conformation or conformations that best satisfy the imposed restraints, and hence also show the lowest scoring function, are sampled. In this matter, most of the data-driven modelling methods can be broadly characterised by their output into two groups; consensus-based modelling and ensemble-based modelling (*Lin, Bonora et al. 2019*). Consensus-based models analytically identify a single 3D conformation returning a consensus structure that best explains the input interaction data. On the other hand, ensemble-based models comprise many conformations that try to account for the variability of the 3C-based datasets.

5.2.1 Consensus-based modelling

Consensus-based modelling methods focus on the modelling of a unique 3D conformation that fits most of the restraints inferred from the input data. These methods have been applied mainly to Hi-C datasets, and they just need to reconstruct one model per dataset, which reduces a lot the computational time.

miniMDS (*Rieber and Mahony 2017*) for example, uses a multidimensional scaling (MDS) algorithm to transform the input interaction matrix into a distance matrix. It combines modelling at three resolution levels that decrease from local to chromosomal structures in order to reconstruct the whole genome with a minimum amount of time and computational costs. In this way, the method first divides the genome into a set of partitions. It then transforms the interaction matrices of each partition into distance matrices by applying an exponent of -0.25 to non-zero values. These partitions are finally modelled independently and arranged together by using as a guide a lower resolution reconstruction of the whole set. MDS performed better than previously available methods, showing correlations with the input data close to 0.7 in particles smaller than 10,000 loci, and of around 0.5 in bigger ones.

Some methods like **GEM-FISH** (*Abbas, He et al. 2019*) try to obtain more reliable models by combining distance measures from FISH with Hi-C interaction data. Concretely, GEM-FISH first simulates the chromatin organisation at TAD resolution by optimising a cost function that weights the Hi-C and FISH data, and by applying polymer physical properties. Then, it reconstructs the

individual TADs separately by a similar cost function, that uses Hi-C data and the radius of gyration estimated by FISH. Finally, both models are combined by integrating the TADs models centred in the TAD-level resolution models. This process is optimised by gradient descent and aims at reducing the calculated cost functions in order to find the best possible conformation of the model. With this approach, Abbas and colleagues showed that a combination of different experimental interaction and distance data could be beneficial to improve the accuracy of chromatin 3D models.

Other methods like **MDSGA** (*Kapilevich, Seno et al. 2019*) work with a combination of graph shortest path algorithms, genetic algorithms, and MDS. Specifically, MDSGA converts the input interaction matrices into distance matrices (known distances) and uses a shortest path algorithm to calculate missing long-range distance data (calculated distances). Then a population of distance matrices are created by modifying the calculated distances based on a defined distribution, and a series of steps of model scoring, merging and mutation (by a genetic algorithm) are applied to obtain the final distance matrix. This matrix will be the one in which the known distances are more similar to the ones calculated from the input data.

These three examples are just a few from the many available methods (*Oluwadare, Highsmith et al. 2019*) and have shown to be useful to model whole-genome structures. However, it is important to note that since most of these methods use population-based interaction data as input, they assume that a single conformation can explain the 3D organisation of the chromatin in the population. In contrast, the experimental data suggest that this not right at the long nor the short genomic scale level (*Nagano, Lubling et al. 2013*, *Bintu, Mateo et al. 2018*, *Nir, Farabella et al. 2018*).

5.2.2 Ensemble-based modelling

Ensemble-based modelling focuses on the modelling of an ensemble of 3D structures in an attempt to reproduce the structural variability found in a population of cells. These methods have also been applied primarily in Hi-C datasets. However, since the modelling step can be quite CPU intensive, most of them focus on

the detailed analysis of specific regions or chromosomes of interest, with some exceptions. The followed approaches diverge depending on the method and its initial focus.

TADbit (*Serra, Bau et al. 2017*) for example, using only Hi-C data as input was able to reconstruct models with enough detail (at the kilobase scale) to detect distinct 3D organisations associated with previously defined specific epigenetic states (*Filion, van Bemmel et al. 2010*). TADbit serves of the Integrative Modeling Platform (IMP) (*Russel, Lasker et al. 2012*) for the application of spatial restraints to a 3D model of the chromatin. For that, Serra and colleagues first normalised the input interaction data and treated it by applying a $\log_{10}$ and a Z-score transformation. Then, they represented the chromatin as a chain of particles with a diameter defined by the resolution of the data. Once all set, they used a combination of parameters to transform the Z-scores of non-consecutive particles into three types of restraints that aimed at placing each pair of particles into a range of allowed distances while consecutive particles were spatially restrained by their occupancy. Finally, the restraints were applied starting from randomly distributed particles and fitted by a series of rounds of Monte Carlo combined with standard simulated annealing. The output of this process is an ensemble of models that best fit the input restraints and minimise the defined scoring function for the given parameter combinations. These ensembles are subsequently compared with the input interaction matrix, and the parameters are optimised to best match the input data.

Hierarchical3DGenome (*Trieu, Oluwadare et al. 2019*), is one of the few ensemble-based modelling methods which aims at high-resolution whole-genome reconstructions, reaching the 5 kb resolution. For this, they split the genome into domains by the arrowhead domain algorithm (*Rao, Huntley et al. 2014*), reconstructing at low resolution the relative position between large scale domains, and then increasing the resolution to resolve the organisation within the domains. In more detail, they first normalise the individual domains interaction data at 5 kb resolution with KR method (*Knight and Ruiz 2012*), and the entire chromosome interaction data at domain resolution with ICE (*Imakaev, Fudenberg et al. 2012*). Then they convert the normalised interaction frequencies into spatial distances by a specific function.

These distances are used to reconstruct both resolution matrices with LorDG (*Trieu and Cheng 2017*). After scaling of the domain resolution models, the method places the centres of mass of the high-resolution models in their corresponding locations. Finally, further optimisation steps are used to fit the inter-domain distances. This process achieves high correlations with the input data.

As a completely different approach, Genomic organisation reconstructor based on conformational Energy and Manifold learning (**GEM**) (*Zhu, Deng et al. 2018*), instead of transforming interaction frequencies into restraints, embeds neighbouring affinities from the interactions into a 3D Euclidean space. To do so, it first uses normalised interaction frequencies as edges to build an interaction network connecting the different genomic bins (nodes) with interaction data. The edges are further optimised by minimising the conformational energy and the Kullback-Leibler (*Kullback and Leibler 1951*) divergence between the inferred 3D Euclidean space and the Hi-C data. This optimisation first retrieves an average conformation of the modelled chromatin, and after a multi-conformation optimisation gives rise to an ensemble of conformations. In this way they have modelled the human chromosome 14 at a 1 MB resolution, obtaining a correlation above 0.9 with the original Hi-C data, and concordance with distances measured by FISH. Furthermore, the method allows the inference of a latent function between the input Hi-C and the output distance data, which is useful to compute the interaction frequency values not present in the input interaction data.

The previous methods mainly focus on the modelling of dense interaction matrices, where most of the possible interactions have frequency data. However, experiments like 4C, single-cell Hi-C, pcHi-C, or HiChIP result in sparse interaction matrices missing most of the possible interactions within the surrounding loci. Hence, to recover the 3D organisation of the defined region, their data has to be treated and modelled in a specific way. In this matter, the number of available methods decreases a lot.

Among these methods, single-cell Hi-C is the one with more alternatives. As an example, Single-cell lattice (**SCL**) (*Zhu and*

Wang 2019) is used to build models of whole chromosomes. In SCL, interactions are binarised by presence or absence of interaction data, and the resulting matrix is transformed into a propensity contact matrix. Specifically, a Gaussian function is used to estimate the propensity of contacts between loci with no interaction data by their linear proximity to other loci with interaction data. Then, the modelling process represents chromosomes as chains of beads with a size proportional to the resolution of the experiment, and places them in cubic cells inside of a 3D cubic lattice. While the modelling process tries to satisfy the restraints from the propensity contact matrix, particles are only allowed to move to their neighbouring cell in each step, reducing the available coordinates in the space, limiting particles movement range, and hence, saving computational time. The final conformation is obtained following an optimisation by simulated annealing to minimise the loss function. With this method, they managed to get twice as correlation with the original data than a previously developed approach.

On the other hand, **4Cin** (*Irastorza-Azcarate, Acemel et al. 2018*) is one of the few methods focused on modelling of sparse capture 3C-based datasets. It uses an approach similar to TADbit but has been optimised to use as input a conjunction of 4C-seq datasets that are treated and combined to obtain sparse interaction matrices. Specifically, the read counts of a minimum of four 4C-seq datasets per modelled region are scaled to the experiment with the biggest number of read counts. Then, interaction data is binned to the size of the produced 4C-seq DNA fragments, transformed into $\log_{10}$, and into Z-scores as in TADbit. The chromatin is then represented as a string of beads with a size proportional to the binned DNA segments lengths. Once all set, a combination of parameters are used to define a viewpoint-specific reach window, in which harmonic distance restraints are applied proportionally to the absolute Z-score values. Outside the reach-window, particles are restrained with harmonic Lower-bound restraints. Jointly, harmonic upper-bound distances are used to restraint consecutive particles. As in TADbit, these restraints are applied starting from randomly distributed particles and fitted by a series of rounds of Monte Carlo combined with standard simulated annealing. The output of this process is an ensemble of models that best fit the input restraints and minimise the defined scoring function. These ensembles of

models can be used to compute a virtual Hi-C interaction matrix from it, and the method has proved to be accurate enough to detect TAD pattern modifications driven by a mutation.

Overall, this picture shows that modelling strategies have focused their attention on technologies like Hi-C, which have been wildly used in the last decade. However, the increase in the usage of other chromatin interaction technologies, like 4C-seq, pcHi-C or HiChIP, that produce sparse interaction datasets requires a step forward for the design of methods suited for them.

CHAPTER 1

3D reconstruction of genomic regions from sparse interaction data

Julen Mendieta-Esteban, Marco Di Stefano, David Castillo, Irene Farabella, and Marc A Marti-Renom. **3D reconstruction of genomic regions from sparse interaction data**. bioRxiv. October 11, 2020

3D reconstruction of genomic regions from sparse interaction data

Julen Mendieta-Esteban[1], Marco Di Stefano[1], David Castillo[1], Irene Farabella[1,*], and Marc A Marti-Renom[1,2,3,4,*]

[1]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.
[2]Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.
[3]Universitat Pompeu Fabra (UPF), Barcelona, Spain.
[4]ICREA, Barcelona, Spain.
*To whom correspondence should be addressed. Emails: *martirenom@cnag.crg.eu* & *irene.farabella@cnag.crg.eu*

ABSTRACT

Chromosome Conformation Capture (3C) technologies measure the interaction frequency between pairs of chromatin regions within the nucleus in a cell or a population of cells. Some of these 3C technologies retrieve interactions involving non-contiguous sets of loci, resulting in sparse interaction matrices. One of such 3C technologies is Promoter Capture Hi-C (pcHi-C) that is tailored to probe only interactions involving gene promoters. As such, pcHi-C provides sparse interaction matrices that are suitable to characterise short- and long-range enhancer-promoter interactions. Here, we introduce a new method to reconstruct the chromatin structural (3D) organisation from sparse 3C-based datasets such as pcHi-C. Our method allows for data normalisation, detection of significant interactions, and reconstruction of the full 3D organisation of the genomic region despite of the data sparseness. Specifically, it produces reliable reconstructions, in line with the ones obtained from dense interaction matrices, with as low as the 2-3% of the data from the matrix. Furthermore, the method is sensitive enough to detect cell-type-specific 3D organisational features such as the formation of different networks of active gene communities.

INTRODUCTION

Chromatin within the nucleus is organised into higher order structures that emerge at different genomic scales, from chromosome territories (at tens of megabases scale), active and inactive chromatin domains (at few megabases scale) [1], self-interacting domains or TADs (at hundreds of kilobases scale) [2-4], and long-range chromatin loops between regulatory elements (at tens of kilobases scale). This multi-scale organization has a direct impact on many biological processes such as gene regulation, DNA replication, and cell differentiation [5-7]. Indeed, genome structure typically reflects cell-type-specific differences in the transcription pattern, and it is frequently rewired upon cell state changes and disease onset [8]. Thus, investigating the principles shaping chromosome three-dimensional (3D) structure is pivotal to shed light into the relationship between genome structure and function.

Several experimental techniques are available to examine chromatin organisation [9]. Among them, molecular biology methods such as Chromatin Conformation Capture (3C) and its derivatives are widely used [10]. These experiments retrieve information about the frequency of interaction between loci in single [11-13] or in populations of thousands to millions of cells and have been designed to analyse the chromatin landscape at different genomic scales [1, 14-16]. For example, some cell population-based experiments allow the retrieval of unspecified interactions in the whole genome (e.g., Hi-C [1], Micro-C [14], GAM [15], and SPRITE [16]). Complementarily, other 3C-based experiments are tailored to capture interactions centred on a specific locus with the rest of the genome (e.g., 4C [17] and multi-contact 4C (MC-4C) [18]) or on sets of dispersed loci in the genome, such as loci enriched for a specific protein (HiChIP) [19] or loci harbouring gene promoters (pcHi-C) [20]. Each class of 3C-based experiments provide different but complementary insights on particular aspects of the genome organization, and their analysis is dependent on the experimental genomic resolution and on the inherent technical biases of each experimental procedures.

A variety of physics- and data-driven approaches for genome 3D reconstruction have been developed to expose the principles shaping chromosome 3D structure [21-24]. For instance, data-

driven (restraint-based) modelling approaches as PSG [25, 26], TADbit [27], 4Cin [28], and TADdyn [29] have been implemented to reconstruct ensembles of chromatin 3D models from cell population-based datasets. Others are focused on the 3D modelling of chromatin based on single-cell Hi-C data, like manifold based optimization [30] and NucDynamics [31]. However, the majority of the data-driven methods are based on interaction experiments that have been designed to retrieve dense contact information from a continuous set of loci or the whole genome, while other interaction experiments are characterised by data sparseness (e.g., HiChIP or pcHi-C). As such, data-driven methods for sparse data modelling are needed.

Generally, the interaction profiles of sparse 3C-based datasets have specific properties that set them apart from other 3C-like techniques characterised by a dense interaction profile. Indeed, protein or promoter capture-based interaction profiles are heavily biased on interactions between captured fragments and devoid of interactions between non-captured fragments. This fact poses the question of whether this lack of information prevents the 3D reconstruction of the whole loci of interest and its analysis, or whether it is sufficient to allow for accurate 3D modelling. To answer this question, we have implemented a new method, which is tailored to integrative modelling and analysis of sparse 3C-based datasets. We have also validated the procedure comparing the resulting reconstructed models with available dense experimental datasets, unveiling that the 3D chromatin organisation can be well recovered by interrogating only a small percentage of loci. Additionally, we have designed new tools to facilitate a robust differential analysis of the resulting models and showcased their usability in comparative analyses using the β-globin locus as a test case. Interestingly, comparing different cell-types, we unveiled that the β-globin locus in cord-blood Erythroblasts (cb-Ery), where its foetal and adult β-globin genes are highly expressed, is hierarchically organised in a 3D network of active gene communities that follows an expression gradient.

RESULTS

Overall modelling strategy for sparse 3C data

Sparse 3C datasets provide information of interactions that involve a limited number of specific loci in the genome. pcHi-C, for example, provides a promoter-centred view of chromatin interactions, helping to assign distal regulatory regions to their target genes, thus providing insights on how gene expression might be controlled [32-34] and how disease-associated genomic variation could affect gene regulation [35]. The main limitation of these sparse technologies, however, is the scarcity of specialized tools for their analysis. Here we have developed an integrative 3D modelling method capable of dealing with data sparsity, enabling the analysis and interpretation of pcHi-C data, and tested it on 12 distinct loci (Benchmarking datasets; **Methods** and **Supplementary Table 1)**. Our method follows an integrative modelling procedure comprising five steps [36]: (i) gather experimental data and process them to obtain the input interaction matrix for the modelling approach, (ii) represent the selected chromatin regions using a bead-spring polymer model with a particle size proportional to the genomic resolution of the experimental data, (iii) transform the frequency of interactions into spatial retrains, (iv) sample the conformational space by steered molecular dynamics, and (v) analyse and validate the obtained ensemble of 3D models (**Methods** and **Figure 1A**).

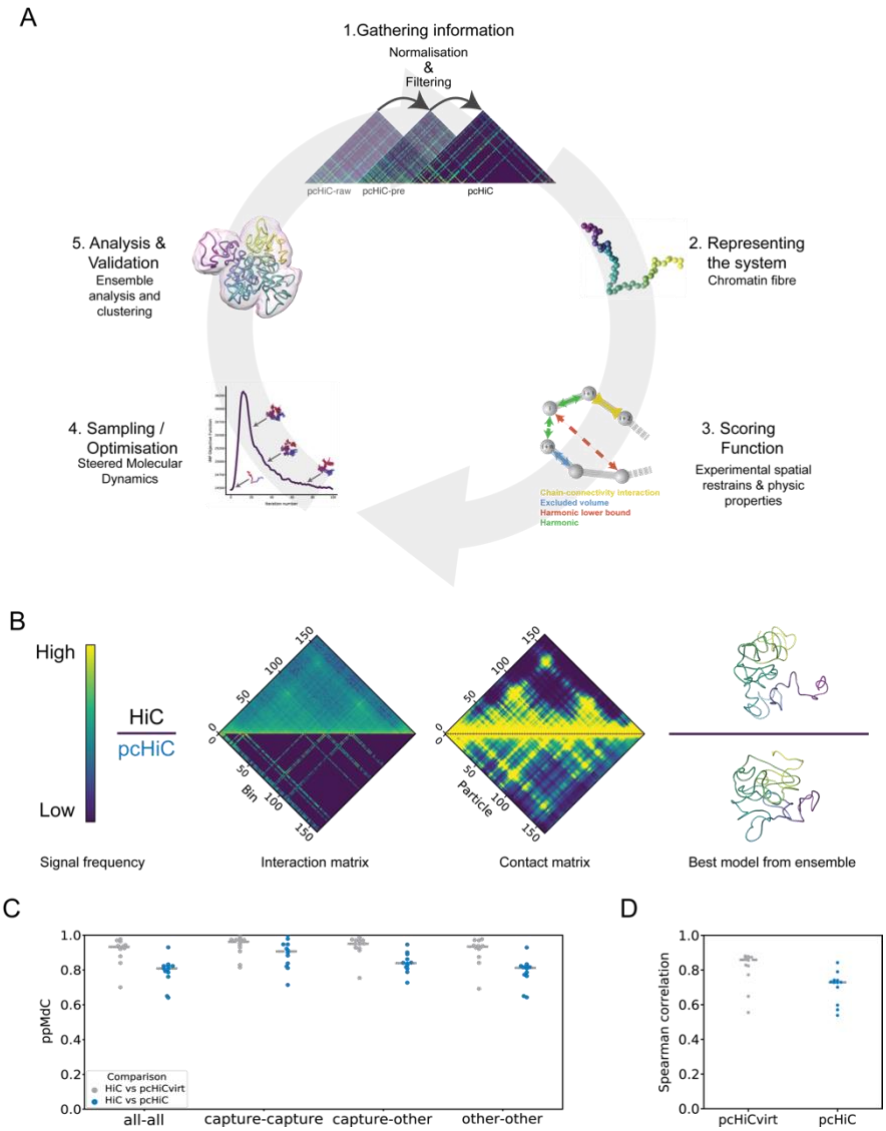


Figure 1. Integrative modelling for sparse datasets efficiently reconstructs the 3D organisation of genomic loci. (A) Workflow of the integrative modelling approach followed to build ensembles of chromatin 3D models from pcHi-C: i) gathering the input interaction matrices with subsequent normalisation and filtering; ii) representation of the chromatin fibre as a polymer with the particle size proportional to the resolution of the experiment; iii) definition of the scoring function used in the modelling procedure. Here, the scoring function comprises spatial restraints derived directly from the input interaction data and from properties of the chromatin fibre (**Method**); iv) sampling the conformational space by steered molecular dynamics (**Method**); and v) validation of the obtained ensemble of

models and further analysis. Model images in all panels were created with Chimera [73]. **(B)** Representation of the input and output data from region 2 (**Supplementary table 1**). The upper half of the panel refer to the dense dataset (Hi-C), whereas the lower half refer to the sparse-datasets (pcHi-C). From left to right, the matrices of normalised interaction frequency (**Methods**) between each pair of bins, the contact matrix obtained from the ensemble of models of region 2 displays the percentage of models in which two bins are found bellow the defined distance cut-off for the contact (**Methods**), and the best model from the ensemble as assesses by the scoring function. The colour bar shows the colour coding from low (blue) to high (yellow) interaction or contact frequencies signal. (**C)** Comparison between models ensembles derived from sparse (pcHi-Cvirt and pcHi-C in grey and blue, respectively) and dense (Hi-C) datasets assessed by the particle-to-particle median distance correlation (ppMdC; **Methods**). Three subsets of particles have been compared given the enclosed loci: (i) captured loci (capture), (ii) non-captured loci (other), and (iii) all the loci (all). The grey dashed line indicates the median ppMdC in the 12 analysed regions. **(D)** Element-wise Spearman correlation coefficients between the experimental Hi-C interaction matrices and the contact maps derived from the model ensembles reconstructed from sparse data (pcHi-Cvirt and pcHi-C in grey and blue, respectively). The grey dashed line indicates the median element-wise Spearman correlation coefficients in the 12 regions analysed.

In this work, we gathered pcHi-C interaction data (**Methods**), whose processing step is pivotal to minimize the experimental biases from the capture protocol. To this end, we designed a multi-stage normalisation procedure named PRoportion of INTeraction approach (PRINT, **Methods**). PRINT weighs each interaction by dividing it by the cumulative whole-genome interaction frequencies of both of the interacting bins, regularising the interaction patterns for the fact that captured loci are highly enriched in contacts. It also removes the pcHi-C unspecific interactions between non-probed bins. To test quantitatively the performance of our normalisation procedure, we compared each of the normalisation stages of the pcHi-C matrices with the respective Hi-C matrices normalised with OneD in each of the selected loci [37]. The median correlation between bins with interaction data in both matrices was 0.27 (+/- 0.025 Median Absolute Deviation (MAD)) for raw pcHi-C matrices (pcHi-C-raw), increasing to 0.44 (+/- 0.032 MAD) with the pcHi-C pre-normalisation step (pcHi-C-pre), and reaching 0.60 (+/- 0.056 MAD) for fully normalised pcHi-C matrices (pcHi-C-norm) (**Supplementary Figure 1A**), suggesting that PRINT reduced successfully the target biases. Then, we represented the selected loci as a bead-spring polymer model with a particle size set to 5 kb, taking into account the restriction fragment lengths distribution in

the benchmarking datasets (**Supplementary Figure 1B**). Similarly to TADbit [27] and TADdyn [29], to simulate the structural conformation of genomic loci, we then transformed the interaction frequencies associated with each bin pair into spatial restraints (**Methods**). The latter were then imposed on the model using steered molecular dynamics as sampling method in which the spring constant associated to each restraint was ramped up as a function of simulation time from zero to the value computed from the interaction data. Lastly, we implemented new means for a robust quantitative spatial differential analysis of genomic loci.

Comparison between sparse and dense 3C-derived models

Dense Chromatin Conformation Capture data has been extensively used to reconstruct the 3D organisation of genomic loci [25, 27, 29, 30]. Here, to test the reliability of our modelling approach, we used sparse and dense datasets to build ensembles of models of the same loci. Specifically, we applied our integrative method for sparse data modelling to previously published pcHi-C datasets of GM12878 cells [32] to reconstruct 3D model ensembles of 12 distinct loci (**Figure 1B** and **Supplementary Table 1**) at a 5kb resolution and compared them with the corresponding ones reconstructed using Hi-C [6] at the same genomic resolution. Additionally, to quantify the effect of sparsity in the comparison independently of the experimental protocol biases, we generated virtual pcHi-C (pcHi-Cvirt) interaction matrices from the normalised Hi-C datasets extracting the rows and columns probed in the pcHi-C experiment (**Methods**). These virtual sparse matrices were then used to reconstruct 3D model ensembles of the selected loci.

The comparison between the sparse and dense derived 3D model ensembles revealed that it is possible to recover most of the 3D organisation of the dense dataset in spite of the data sparsity (**Figure 1C**). Indeed, the all-vs-all particle-to-particle median distance correlation (ppMdC) between the sparse and dense derived 3D model ensembles was 0.81 (+/- 0.019 MAD) and 0.93 (+/- 0.024 MAD) for both pcHi-C and pcHi-Cvirt**.** Additionally, when comparing distances between particles that have both been captured in the pcHi-C experiment (capture-capture), the ppMdC was higher, reaching 0.91 (+/- 0.054 MAD) for pcHi-C and 0.96 (+/- 0.019 MAD) for pcHi-Cvirt. Consistently, when comparing distances between non-captured particles with captured particles (capture-

other) or between non-captured particles (other-other), the ppMdC indicated good agreement with values of 0.84 (+/- 0.03 MAD) and 0.95 (+/- 0.02 MAD), and 0.81 (+/- 0.02 MAD) and 0.93 (+/- 0.02 MAD) respectively for pcHi-C and pcHi-Cvirt in both comparisons (**Figure 1C**)**.** The results indicate that the sparse derived ensembles of 3D models are a good representation of the dense experiment and that the intrinsic experimental biases of the capture experiment only minorly affect the 3D reconstruction. Indeed, comparing the whole contact map computed from the 3D model ensembles derived from sparse data directly with the whole experimental Hi-C interaction matrices revealed that the reconstructed ensembles of models are in good agreement with the dense experimental data having an element-wise Spearman's rank correlation coefficient of 0.73 (+/- 0.02 MAD) and 0.86 (+/- 0.02 MAD), for pcHi-C and pcHi-Cvirt derived ensembles of models, respectively (**Figure 1D**). Overall, this suggest that the ensembles of models reconstructed by our approach represent well the 3D organisation of the selected genomic regions and, more importantly, recover the spatial arrangements of loci that are not interrogated by the sparse experiment.

Reconstruction efficiency and data sparsity

To investigate the relationship between the reconstruction efficiency and data sparsity, we simulated 'synthetic' capture data. Briefly, we generated 10 different sets of 'synthetic' capture matrices that represent generic capture-like experiments. We started from the contact matrix derived from a 3D toy-genome models ensemble that simulates roughly a one Mb length genome (comprising more than 600 particles) with a TAD-like architecture, a high level of interaction noise, and low variability between models [38] (**Methods** and **Figure 2A**). To build each of the 10 'synthetic' sets, we randomly selected 22 captured loci and constructed 6 additional datasets of different sparsity down-sampling each set considering 2, 4, 6, 10, 14, and 18 loci at a time, which mimics the distribution of captured probes per Mb present in a typical genome-wide pcHi-C experiment (**Figure 2B**). The constructed 70 capture-like matrices thus aim to represent typical pcHi-C experimental design. Using our integrative modelling method for sparse datasets, we reconstructed, from each of the 'synthetic' capture matrices in the dataset and their down-sampled counterparts, ensembles of 100 models, and compared them with

the reference toy-genome ensemble (**Figure 2A**). Independently of the sets, the ppMdC between the sparse and dense model ensembles increased with the number of captured particles used in the modelling procedure reaching a median correlation between sets of 0.82 (+/- 0.02 MAD with just 10 captures per Mb (**Figure 2C**). Notably, also with 4 and 6 captures per Mb the ppMdC reached 0.69 (+/- 0.04 MAD) and 0.79 (+/-0.05 MAD) for 4 and 6 captures, respectively, although with greater variation within sets. This suggests that with 10 captured loci per Mb the uncertainty in the input information is smaller, leading to more precisely reconstructed models. Nevertheless, it is possible to reconstruct good models also with fewer as 4 captured loci per Mb although with a higher degree of variability. To quantify the effect of data sparseness on model reconstruction, we next measured the amount of input information used during the modelling as the percentage of all possible interaction pairs in the contact matrix (dense data input) and then assessed it with the ppMdC. The results indicate that it was possible for the majority of the sets (8/10) to reliably reconstruct the reference toy genome (ppMdC > 0.8) with just 2-3% of all the interaction pairs in the contact matrix used as restraints (**Figure 2D**)**.** Taken together, this analysis shows that it is possible to consistently recover most of the 3D organisation of a region of interest with 10 captured loci per Mb and with just 2-3% of all possible interactions within a region captured.

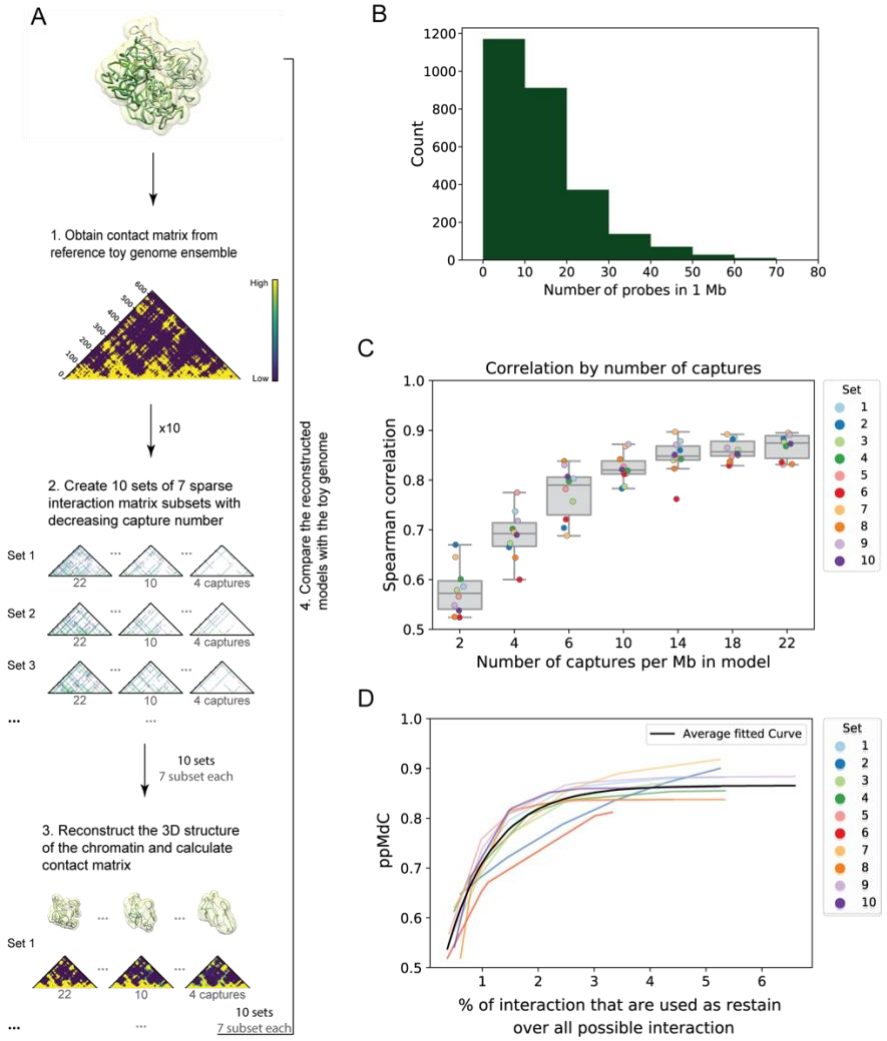


Figure 2. A low percentage of the interaction data is needed to produce reliable 3D reconstructions. (A) Workflow for the generation of 3D model ensembles from 'synthetic' sparse datasets and comparison with the toy genome. A total of 70 'synthetic' captured map were generated representing 10 different capture experiments with different level of data sparsity (**Methods**). Model images were created with Chimera [73]. **(B)** Distribution of pcHi-C probes per megabase windows in the genome [32]. **(C)** Distribution of the ppMdC between the 'synthetic' models and the toy genome grouped by subsets of captures per megabase. Box boundaries represent 1st and 3rd quartiles, middle line represents median, and whiskers extend to 1.5 times the interquartile range. The ten sets of captured positions are displayed with the colour code shown in the insert. **(D)** Relationship between the ppMdC and the percentage of cells in the matrix used as restraints in each set represented with an exponential fit. The used colour code is the same as in **C**, the grey line represents the mean fit of all the datasets in analysis.

Cell-type specific organisation of the β-globin locus

To illustrate the utility of our integrative approach in unveiling the differential organisation of loci, we applied it to the genomic region surrounding the β-globin locus in 3 different cell types for which pcHi-C data are available [33], namely cord-blood derived Erythroblasts (cb-Ery), naive CD4+ T-cells (nCD4), and Monocytes (Mon). The selected genomic region contains five coding genes (HBB, HBD, HBG1, HBG2, and HBE1) with developmental-stage-dependent expression [39], which is finely regulated by a set of upstream enhancers known as the Locus Control Region (LCR) [40]. This locus is known to be in an active conformation in cb-Ery, where the LCR is interacting mainly with expressed genes as HBB and HBD, but not in nCD4 and Mon cells [33].

First, we defined the optimal region to be modelled based on the interaction networks (in all cell types) of the embryonic (HBG1 and HBG2) and adult (HBB and HBD) globin genes with the rest of the genome at 5 kb resolution (**Methods**). The defined region spanned 4.7 Mb of chr11 (chr11:3,795,000-8,505,000 base-pairs (bp)) comprising several neighbouring genes and multiple long-range regulatory elements. By applying our integrative approach, we generated an ensemble of 1,000 3D models for each cell type. The packing of the genomic region was significantly different in each cell types with median radius of gyration of 248+/-3, 242+/-2, and 237+/-2 nm for cb-Ery, nCD4 and Mon, respectively (p-values < $9.1e^{-163}$ in each of the pairwise comparisons using two-samples Kolmogorov-Smirnov statistics) (**Supplementary Figure 3A**), with the topology of the region in cb-Ery being less tightly packed than in nCD4 and Mon. Each ensemble was then clustered by structural similarity [27] and the models from the most populated cluster were selected for the comparative analysis between cell-types. Clustering by distance root-mean-square deviation (dRMSD), confirmed that the topology of the region was markedly different in the three cell types, with nCD4 and Mon folds being more similar between each other than with cb-Ery (**Figure 3A**). Particularly interesting is how the topology of the β-globin locus (chr11:5,201,270-5,302,470) varied in the three cell types. Indeed, in Erythroblasts the β-globin locus appeared to be located further from the main core of the region as compared with naïve CD4+ T-cells and Monocytes, with median distances between the centre of mass of the β-globin locus of 286, 243, and 207 nm in cb-Ery, nCD4, and Mon, respectively

(p-values < $3.46e^{-101}$ in all the pairwise cell-type comparisons; two-samples Kolmogorov-Smirnov statistic) (**Supplementary Figure 3B**).

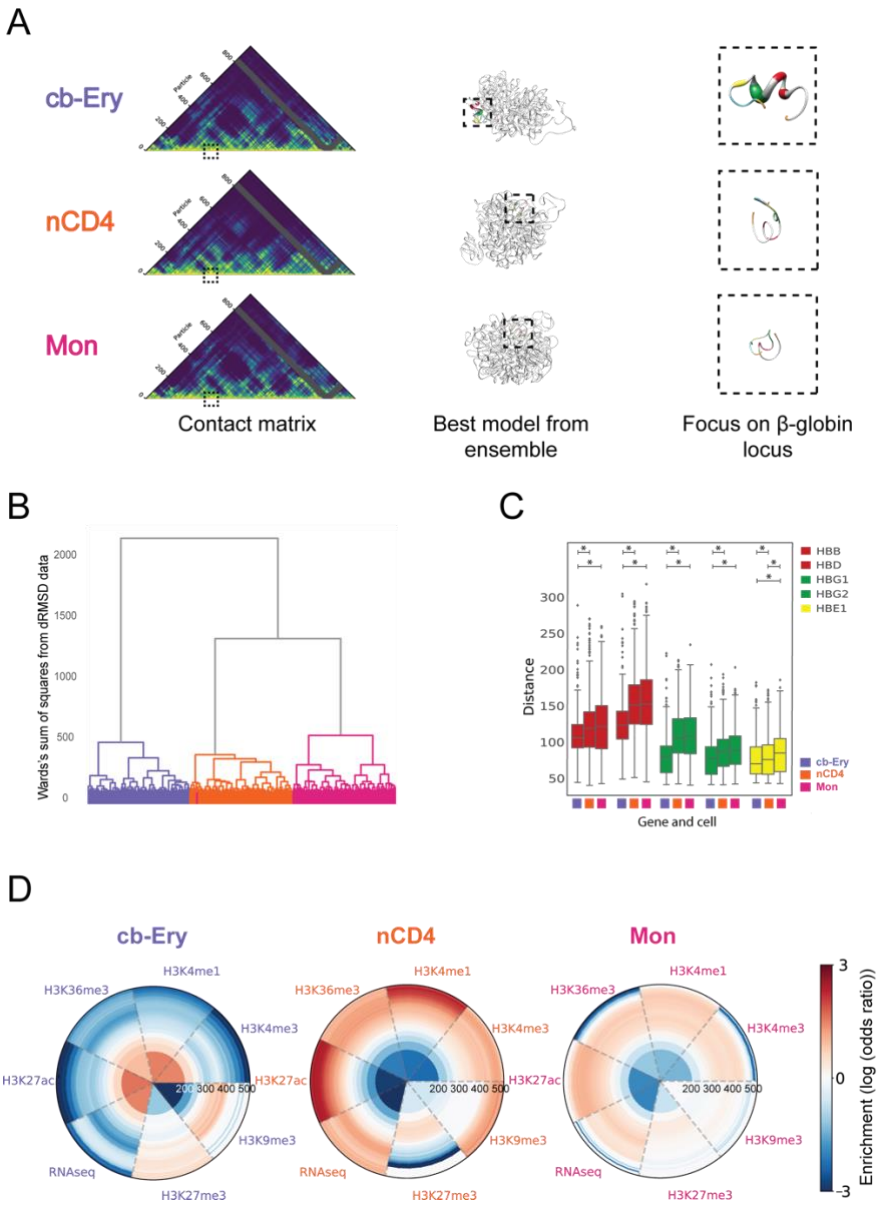


Figure 3. Cell-type specific organisation patterns of the β-globin locus. (A) β-globin locus in cb-Ery, nCD4, and Mon cell-types. From left to right: representation of the contact matrix derived from each of the model ensembles colour coded from low (blue) to high (yellow) contact frequency (columns filtered due to low

interaction data are coloured grey); best model from ensemble as assesses by the scoring function; zoom up of the β-globin locus in the model. Models are represented as a tube with thickness proportional to the cell-type expression profile (**Methods**), the regulatory elements and genes in the β-globin locus are coloured as follow: HBB and HBD in red, HBG1 and HBG2 in green, HBE1 in yellow, LCR in blue and 3'HS1 and HS5 in orange. Model images were created with Chimera [73]. **(B)** Clustering tree (see *Hierarchical clustering of ensembles of 3D models* in Chromatin ensemble 3D analysis) of cb-Ery (purple), nCD4 (orange) and Mon (pink) model ensembles. **(C)** Cell-type specific distance distributions between the particle containing HS3 site of the LCR and the β-globin genes (HBB, HBD, HBG1, HBG2, and HBE1, colour coded as in (A)) as observed in the ensemble of models. Box boundaries represent 1st and 3rd quartiles, middle line represents median, and whiskers extend to 1.5 times the interquartile range (two-samples Kolmogorov-Smirnov test, asterisk indicate $p < 0.007$). **(D)** Radial plot showing the 3D enrichment around HS3 (**Method**). Each circumference shows the enrichment or depletion of features around HS3 on layers (up to 560 nm away from HS3) of non-overlapping volumes equal to the one of the initial sphere with radius of 200 nm. The colour bar shows the colour coding from highly depleted (blue) to highly enriched (red) features.

To characterise this further, we focused specifically on the β-globin locus and quantified its spatial organisation with respect to hypersensitive site 3 (HS3) in the LCR, which is forming an intricate network of interaction with the β-globin genes [41] and is required for their activation [42]. In line with this evidence, in the 3D ensemble of models representing cb-Ery cells, HS3 was significantly closer to HBB, HBD, HBG1, HBG2, and HBE1 genes than in the 3D ensemble of models representing nCD4 and Mon (p-values < 0.007, two-samples Kolmogorov-Smirnov test). In the latter two cell-types HS3 had a similar distance distribution with HBB, HBD, HBG1, and HBG2 genes (p-values > 0.01, two-samples Kolmogorov-Smirnov test) (**Figure 3B**).

Performing 3D enrichment analysis of varied epigenetic features and expression levels around HS3 (**Methods**), we unveiled a stark enrichment of active chromatin marks (H3K27ac, H3K36me, H3K4me1, and H3K4me3) and expression levels, and a clear depletion of inactive marks (H3K9me3 and H3K27me3) in cb-Ery. This 3D functional signature was absent in nCD4 and Mon, where active chromatin marks and transcript levels were depleted (**Figure 3C**). Overall, our models recapitulated the different 3D organisation of the β-globin locus and highlight the existence of a specific 3D

functional signature enriched in active chromatin features that characterised the active β-globin locus in cb-Ery.

Active gene communities in cb-Ery: a cell-type specific 3D signature

To examine whether the specific 3D functional signature of the active β-globin locus influence its genomic neighbourhood, we investigated its long-range interaction patterns. Comparative analysis of the distance profile between HBG2 (the most expressed gene in cb-Ery) and each of the selected loci (chr11: 3,795,000-8,505,000 bp), revealed the existence of an intricate cell-type specific network of spatially proximal expressed genes (**Figure 4A**), in line with previous observations of transcribed genes co-localizing in space [24, 43-46]. This network comprised distal transcribed sites (even located at 1.4 Mb away as STIM1) that showed cell-type specific spatial proximity. Indeed, HBG2 in cb-Ery was in closer proximity with all other expressed loci of the genomic neighbourhood than in nCD4 and Mon (**Figure 4B**).

To further characterise the cell-type specific spatial distribution of these transcribed loci, we clustered their relative distances within the ensembles of 3D models and identified communities of expressed genomic loci (**Figure 4C-E** and **Methods**). Then, we quantified the amount of times a given community of expressed genomic loci occurred within the ensembles of 3D models (*i.e.*, the co-occurrence score, **Methods**) and used this quantification as a proxy to define the "community stability". This analysis revealed the existence of highly variable communities of expressed genomic loci that followed a cell-type specific segregation in the 3D space. Interestingly, the organization of these communities was overall more stable in cb-Ery than in nCD4 and Mon, where less defined communities were identified. Indeed, as assessed by the mean inter-community co-occurrence scores (**Methods**), the cb-Ery network was characterised by the presence of four stable communities (**Methods** and **Table 1**). While, the nCD4 network was formed by three communities with overall low co-occurrence (although community 2 in this network showed a stability in line with the communities in the cb-Ery network), and the Mon network formed by only two unstable communities (**Methods** and **Table 1**). Overall, the results highlight the presence of more defined 3D communities of expressed genes in cb-Ery as compared to nCD4 and Mon, suggesting that the co-occurrence of these segregated communities

within an ensemble of possible folds is part of the cell-type specific 3D signature.

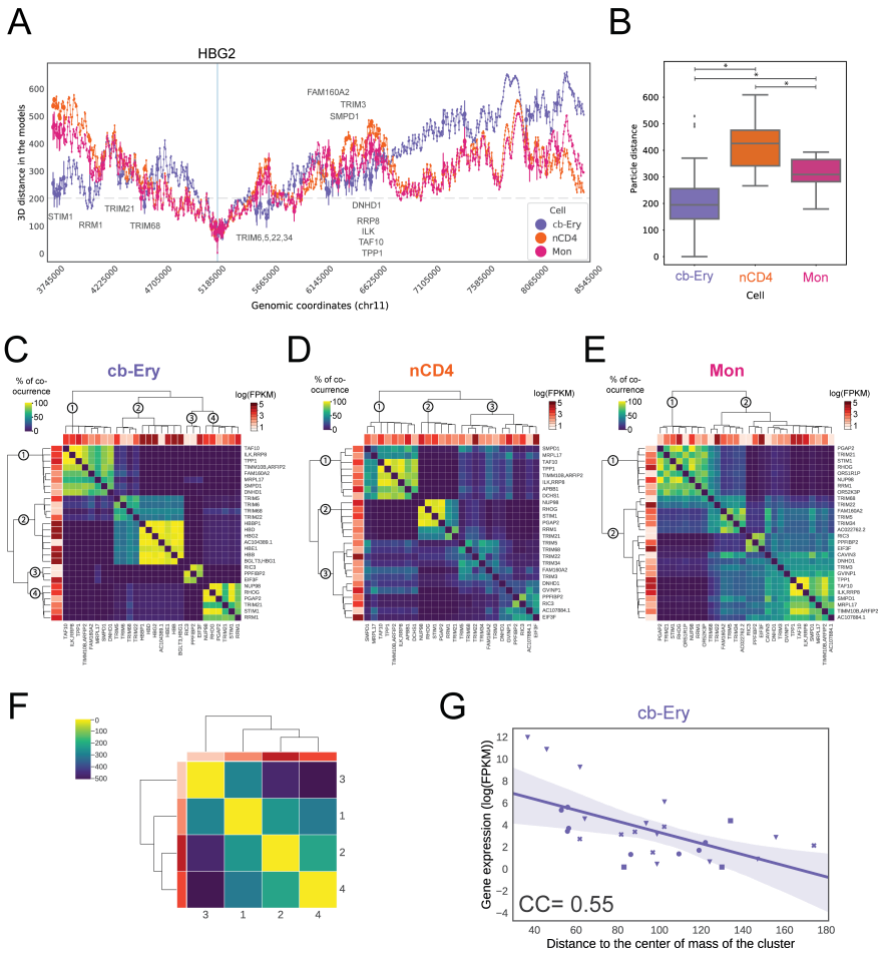


Figure 4. Communities of active genes as a cell-type specific 3D signature in cb-Ery. (A) Line plot of the mean distances between the TSS of HBG2 (focus point, blue vertical line) and all other particles in the genomic region (chr11:3,795,000-8,504,999 bp) for cb-Ery (purple), nCD4 (orange), and Mon (pink) as calculated in each model ensembles. Error bar, indicating one standard deviation, is displayed for particles enclosing a transcribed gene (in at least one cell). The grey dashed line indicates 200 nm cut-off used in the analysis (**Methods**). **(B)** Cell-type specific distance distribution between particles enclosing the HBG2 gene and all transcribed genes in the genomic region (chr11:3,795,000-8,504,999 bp) for cb-Ery (purple), nCD4 (orange), and Mon (pink) as calculated in each model ensembles. Box boundaries represent 1st and 3rd quartiles, middle line represents median, and whiskers extend to 1.5 times the interquartile range (two-samples Kolmogorov-Smirnov test, asterisk indicate p-values < $7.5e^{-6}$). **(C-E)** Hierarchical clustering of each genes based on the co-occurrence analysis (**Methods**) in cb-Ery (**C**), nCD4 (**D**)

and Mon (**E**). Co- occurrence value range from 0 (low, dark blue) to 100 (high, bright yellow). In each hierarchical tree the communities are labelled at their root branch. Per each gene the relative expression (log(FPKM) is shown in a scale of reds from 0 to 5. **(F)** Hierarchical clustering of the distances between the communities defined in cb-Ery (**Methods**). Distance values are coloured in the matrix from dark blue to bright yellow and the average expression in log(FPKM) per community is coloured by ranking from lowest (lightest) to highest (darkest) in 3 different shades of red. **(G)** Relationship between gene expression in log(FPKM) and the median distance of the gene particles to the centre of mass of its own community in cb-Ery ensemble of models (**Methods**). Purple line denotes the linear regression fit, the shading around the regression line represents the confidence interval, each community is represented with different symbols (circle community 1; inverse triangle community 2; square community 3; and ex community 4).

Table 1. Communities stability assessment

Cell	Community	Mean inter-community co-occurrence	Average inter-community co-occurrence per cell
cb-Ery	1	2.96	3.06
	2	4.90	
	3	0.54	
	4	3.85	
nCD4	1	11.49	9.16
	2	3.83	
	3	12.17	
Mon	1	10.33	10.33
	2	10.33	

Description: **Cell**, the cell-type data used to reconstruct the chromatin; **Community**, the defined communities by Ward's clustering; **Mean inter-community co-occurrence**, Communities stability score as defined in **Methods**; and **Average inter-community co-occurrence per cell**, average Mean inter-community co-occurrence value of all the communities in each of the cells.

Next, we investigated whether the stability of the 3D communities of expressed genes in cb-Ery could be related to the high levels of expression of the β-globin genes (highest as HBG2 with 10.86 FPKM, while the mean expression of all the other expressed genes in nCD4 and Mon was 2.45 and 2.10 FPKM respectively). Clustering the distance distribution between the centres of mass of

each community in cb-Ery (**Figure 4F**) revealed a clear hierarchical organisation with the most expressed community, which included the highly expressed β-globin locus (**Supplementary Table 2**), located in the centre, and the least expressed community in the periphery. This pattern was not present in nCD4, and impossible to address in Mon with just two communities (**Supplementary Figure 4A-B**). This suggests a hierarchical organisation in cb-Ery, in which the location in space of each of the communities and their levels of expression are related. Surprisingly, this hierarchy was also overall present at the community level in cb-Ery, where the distance between each gene to the centre of mass of the community and its expression were negatively correlated (CC: -0.55, p-value=0.002; **Figure 4G**). This suggests the formation in cb-Ery of a gradient of expression within the community were the most expressed genes are located in the centre of their communities and the less expressed ones are preferentially located in the periphery in line with the organisation previously observed for the alpha-globin locus [24]. This overall community organisation was not evident in nCD4 and Mon (**Supplementary Figure 4C-D**), thus suggesting that the high expression of the β-globin loci in cb-Ery could be associated with the establishment of a hierarchical organisation in the loci.

DISCUSSION

Here, we have introduced an integrative modelling method for the 3D reconstruction, analysis, and interpretation of sparse 3C-based datasets such as pcHi-C. We also demonstrate its usability in the comparative 3D analysis of genomic regions using the β-globin locus as an example, showing that our method can detect cell-type-specific 3D organisational features within genomic regions that can lead to several important implications on the relationship between genomic function and spatial genome organisation, such as the expression dependent organisation of active loci.

Generally, the analysis and interpretation of sparse 3C-datasets is not trivial and specialised analytical tools are required. In the case of pcHi-C, the available tools (ChiCMaxima, Chicago, Gothic, Chicdiff, HiCapTools [47-51]) are mainly focused on the implementation of normalization strategies to reduce the impact of non-biological biases and on strategies to detect interaction between captured loci. Conversely, the integrative modelling method

presented in this study has been designed for the analysis and interpretation of sparse 3C-datasets in their third dimension, allowing for data normalisation, detection of significative interaction, and most importantly, the recovery of the full structural organization of a genomic region despite of the data sparseness.

Indeed, here we extensively tested our procedure by comparing models reconstructed directly from sparse and dense datasets, showing that 3D models reconstructed by the integrative modelling method for sparse data modelling are a good representation of the dense experiment. In fact, model reconstruction is only minorly affected by the intrinsic experimental biases of the capture experiment. Additionally, and most importantly, our model procedure reproduces remarkably well the whole 3D organisation of the selected genomic regions even recovering the organisation of loci that are not included as input restraints and are not readily observable in the sparse experiment.

Next, to assess whether the 3D reconstructed models were not only a bona fide representation of models based on Hi-C datasets, we used a 'synthetic' toy genome with known 3D organisation [38] and proved that we can efficiently model sparse pcHi-C-like datasets using as few as 2-3% of all possible interaction data. Importantly, this quantification highlights how the degree of sparseness of the data is related to the efficiency of the 3D reconstruction process and provide a general guideline for sparse data modelling. In light of this, we speculate that our integrative approach could easily be applied to different type of 3C datasets with similar sparseness. For example, protein-centric chromatin conformation method such as HiChIP [19] could be used as input experiment to reconstruct the chromatin folding, assuming that the protein-capture biases of this type of experiments are similar to the promoter-capture biases observed in the pcHiC experiments.

Finally, to illustrate the utility of our integrative approach, we applied it to the β-globin locus, whose 3D organisation has been extensively studied [39, 41, 52-54]. We investigated this locus in three different cell types (cb-Ery, nCD4, and Mon) and performed a comparative analysis between them. In agreement with previous studies [33], our models show that the topology of the β-globin locus varies in the three cell types owing to their differential

expression. Interestingly, our models also unveil that the globin HBG2 gene is embedded in an epigenetically active and highly transcribed neighbourhood in cb-Ery giving rise to a locus-specific 3D functional signature. This functional signature is absent in the models of other cell-types (nCD4 and Mon), where the locus is not expressed. We also show that this cell-specific organisation, not only occurs proximally to the β-globin genes but also involves loci located at longer genomic distances (more than 1 Mb away). Indeed, our 3D comparative analysis unveiled the existence of an intricate cell-type specific network of spatially-proximal expressed genes that forms gene communities that are segregated in the 3D space in a cell-type specific fashion. The identified communities are compatible with the formation of chromatin foci in which transcribed genes co-localize as a general mechanism to organise gene transcription [24, 43-46, 55]. Interestingly, we observed that the co-occurrence within the ensemble of models of the identified cell-type specific communities is cell-type dependent, with the cb-Ery communities network formed by more persistent communities than the nCD4 and Mon community networks. This suggests that also the degree of co-occurrence of the communities within the ensemble is an important feature for the identification of a cell-type specific 3D signature. Additionally, we observed that in cb-Ery, where the β-globin genes are highly expressed, the communities present an overall hierarchical spatial organisation, both between and within communities. This topology is dependent on the level of transcription with highly expressed entities (entire community or specific gene within a community) located in the core of the hierarchical 3D organisation and low expressed entities found at the periphery. We hypothesise that the observed communities could represent cell-type specific transcription factories [24, 55-57] or phase-separated foci [58-60] organised following a gradient of transcription with high concentration of nascent transcripts and transcription machinery in the core of the assemblies that create a "sticky" environment to the less expressed peripheral loci. This hierarchical organisation is only marginally present in nCD4 and Mon, suggesting that it also contributes to the cell-type specific 3D signature characterising the β-globin region in cb-Ery.

In summary, we have shown that sparse datasets like pcHi-C can be effectively used to model in 3D the spatial conformation of genomic domains. The resulting models retain most of the genomic region

organization and recover also the organisation of loci that are not readily observable in the sparse experiment. Importantly, this is achievable with a very small percentage (~2-3%) of all possible interaction data in the genomic region. Additionally, our study not only provides a novel approach for sparse-data 3D modelling but also introduces new tools for the comparative analysis of genomic regions. Thus, it will aid the discovery of cell-type specific 3D signatures and help deciphering complex mechanism underlying the cell-type specific 3D genome organization.

METHODS

Experimental datasets

Structural data were obtained from publicly available 3C-based chromatin interaction experiments of GM12878 cells (Hi-C GEO: GSE63525 and pcHi-C ArrayExpress: E-MTAB-2323) [6, 32], and cb-Ery, nCD4, and Mon cells (pcHi-C EGA: EGAS00001001911) [33].

Hi-C datasets processing. The reads for each replicate were mapped onto the GRCh38 reference genome, filtered, and merged using TADbit with default parameters [27]. Then, starting from the merged filtered fragments, the genome-wide raw interaction maps were binned at 5 kilo-base (kb) and normalized using OneD [37] as implemented in TADbit [27].

pcHi-C datasets processing. For each experiment, the reads were mapped onto the GRCh38 reference genome using TADbit [27] and were filtered applying the following filters: (i) self-circles, (ii) dangling-ends, (iii) errors, (iv) extra dangling-ends, (v) duplicated reads, and (vi) random breaks. Next, we computed the reproducibility score to measure the similarity between replicates from each pcHi-C dataset [61]. Then, for each cell-type, the different replicates from the same experiment were merged into one dataset for further analysis, making an exception with replicate ERR436029 from the GM12878 pcHi-C dataset (E-MTAB-2323), which was discarded due to a clearly low reproducibility score when compared with the rest of the replicates (average of 0.24 with the other replicates as compared to the average of 0.84 obtained between the other replicates). Using the merged filtered fragments, the genome-wide raw interaction maps of each cell-type were

binned at 5 kb and normalised using the PRoportion of INTeraction approach (PRINT, next section).

Sparse data normalization: PRoportion of INTeraction approach (PRINT). PRINT, a multi-stage normalisation procedure, weighs each pair of interacting bins with the same philosophy as the visibility approach for Hi-C [62]. Starting from a raw interaction matrix as input, PRINT first transforms the raw interaction between two bins (*i* and *j*) into a percentage of interaction with respect to the rest of the genome as:

$$value_{ij} = \frac{bin_{ij}}{\sum row_i + \sum row_j - bin_{ij}}$$

where (bin_{ij}) represent the number of times in which bin *i* and *j* interact, and $\sum row_i$ and $\sum row_j$ are the sum of all the interactions of bins *i* and *j* respectively with all the genome (self-interactions included). Then, the non-baited interactions (that is, those bins containing only pcHi-C off-target reads) are filtered out.

PRINT assessment. Using the *benchmarking datasets* described above, each stage of PRINT normalisation (pcHi-C-raw, pcHi-C-pre and pcHi-C-norm) was assessed in comparison with the dense Hi-C interaction matrix by calculating the Spearman's rank correlation coefficient between interactions (bin_{ij}) present in both interaction matrices.

Reconstructed 3D genomic regions

Benchmarking datasets. We selected 12 genomic regions of interest (**Supplementary Table 1**) as defined by Rao and colleagues [6]. This set of genomic regions were predicted to result in reliable 3D models based on their > 0.7 MMP scores [38] (**Supplementary Table 3**). Briefly, MMP score takes into account the interaction matrix size, the contribution of significant eigenvectors in the matrix, and the skewness and kurtosis of the z-scores distribution of the matrix to assess their potential for being modelled [38].

Comparative analysis datasets. We selected a genomic region around a locus of interest (here the β-globin) defining it in a semi-automatic manner in each cell type. Briefly, a viewpoint, which may be constituted by a bin or a set of bins of interest, is selected.

Here, as viewpoint we used bins enclosing the active haemoglobin genes in cb-Ery (HBB, HBD, HBG1, and HBG2). Then, all the other bins that interacted with the viewpoint bins in the normalised genome-wide interaction matrix were selected. Each of these bins were then scored by their cumulative normalised interaction frequency values with the viewpoint bins. From this set only the top intra-chromosomal 200 bins were selected since, by visual inspection, they were the bins spanning the genomic region that best enclosed the viewpoint. Then an unweighted interaction network was generated with the nodes corresponding to the top 200 bins and the viewpoint bins. Edges between nodes were added if their pairwise cumulative normalised interaction frequency value was in the top 200 interacting bins. Then, a series of transformations were applied to the unweighted interaction network: (i) nodes that are highly proximal in 1D genomic resolution (closer than 25 kb) were merged into one node; and (ii) poorly connected nodes in the network that had less than 5 edges were filtered out (average number of edges per node in Mon, nCD4, and cb-Ery were 200, 214, and 214, respectively). The extreme nodes in terms of genomic coordinates were selected from the final unweighted interaction network to represent the optimal genomic region around the viewpoint. Here, to perform comparative analysis, we defined the optimal genomic region around the viewpoint as the broader genomic region that enclosed all of the genomic coordinates identified in each cell-type.

3D chromosome ensemble reconstruction from sparse datasets

Model representation. Each genomic region was described with a beads-on-string model based-on the previously implemented protocols [29, 63] without bending rigidity potential. Thus, a chromosome was represented with N spherical beads with diameter $\sigma = 50$ nm that contain 5 kb of chromatin which determined the genomic unit length of each model.

System set up for molecular dynamics. All simulations were done using TADdyn [29]. A generic random self-avoiding walk algorithm was used to define the initial conformation of each model. The potential energy of each system comprised the terms of the Kremer-and-Grest polymer model [64] including chain-connectivity (Finitely Extensible Nonlinear Elastic, FENE) [65] and excluded volume (purely repulsive Lennard-Jones) interactions. The

initial conformation was placed randomly inside a cubic simulation box of size 1,000 σ centred at the origin of the Cartesian axis O = (0.0, 0.0, 0.0), tethered at the centre of the box using a harmonic (K_t=50.0 k_BT/σ^2 and d_{eq}=0.0 σ) to avoid any border effect and energy minimized using a short run of the Polak-Ribiere version of the conjugate gradient algorithm [66] to favour smooth adaptations of the implementations of the excluded volume and chain connectivity interaction.

Encoding sparse data into TADdyn restraints. TADdyn [29] empirically identifies the three optimal parameters to be used for modelling based on a grid search approach. This are: (1) maximal distance between two non-interacting particles (*maxdist*); (2) a lower-bound cut-off to define particles that do not frequently interact (*lowfreq*); and (3) an upper-bound cut-off to define particles that frequently interact (*upfreq*). All possible combinations of the parameters were explored in the intervals *lowfreq* = (-1.0,-0.5, 0, 0.5), *upfreq* = (-1, -0.5, 0, 0.5), *maxdist* = (200, 300, 400, 500) nm, and assessing each combination using distance thresholds to determine if two particles are in contact (*dcutoff*) at 100,150, 200, 250, 300, 350, 450, 500 nm. For each of the combinations an ensemble of 100 3D models was generated and the Spearman correlation coefficient between the contact map derived from each ensemble and the experimental input interaction matrix was calculated. The top set of parameters for each region in each cell-type were set for those resulting in the highest Spearman correlation coefficient between the models contact map and the input interaction matrix. To allow for a robust comparative analysis (**Methods**) the optimal *maxdist* and the *dcutoff* parameters were selected based on the consensus within the top optimal values for each region in each cell-type. Optimal *maxdist* and the *dcutoff* were set at 300 nm and 200 nm, respectively for the ensembles of models reconstructed from the GM12878, cb-Ery, nCD4, and Mon pcHi-C datasets. Once the three optimal parameters were defined, the type of restraints between each pair of particles was set considering an inverse relationship between the frequencies of interactions of the contact map and the corresponding spatial distances. Non-consecutive particles with contact frequencies above the upper-bound cut-off were restrained by a harmonic oscillator at an equilibrium distance, while those below the lower-bound cut-off were maintained further apart than an equilibrium distance by a

lower-bound harmonic oscillator. To identify 3D models that best satisfy all the imposed restraints, the optimization procedure was then performed using a steered molecular dynamic protocol. A total of 1,000 replicate trajectories were generated for each genomic region and dataset. Per each of the 1,000 replicate trajectories, the conformation at the end of the steering protocol (when the target spring constant and equilibrium distance are reached) was retained to form the final ensemble of 1,000 3D models. For the cb-Ery, nCD4, and Mon datasets, to account for possible mirrored 3D models within the final ensemble of 3D models, each ensemble was then clustered based on structural similarity score as implemented in TADbit [27] and only the models from the most populated cluster were retained for further analysis.

Steered Molecular Dynamics protocol. A steered molecular dynamics protocol was used to progressively favour the imposition of the defined set of restraints between non-consecutive particles. For each restraint, the equilibrium distance was set to 1 particle diameter (σ). The spring constant *k(L,t)* was weighted with the sequence-separation *L* between the constrained beads as in TADdyn [29] to ensure that the steering process was not dominated by the target pairs at the largest sequence separation. However, here the *k(L,t)* was smoothly ramped during the steering phase from zero to its maximum value.

3D chromosome ensemble reconstruction from dense datasets

The reconstruction of 3D models of genomic regions from dense data followed the modelling protocol described above. That is, a grid search approach was used to select for the optimal parameters to be used for modelling. The optimal *maxdist* and the *dcutoff* parameters were selected based on the consensus within the top optimal values for each region in the GM12878 pcHi-C dataset and set at 300 and 200 nm, respectively. Using these parameters, the final ensemble of 1,000 3D models was obtained starting from the computed 1,000 steered molecular dynamics trajectories.

3D chromosome ensemble reconstruction from Virtual pcHi-C derived from dense datasets

A dataset of Virtual pcHi-C interaction matrices was produced starting from the normalised Hi-C interaction matrices at 5kb resolution (GM12878 cells GEO: GSE63525; **Methods**) and from

the liftover (*https://genome.ucsc.edu/cgi-bin/hgLiftOver*) list of captured fragments in pcHi-C GM12878 experiment [32]. The obtained Virtual pcHi-C interaction matrices comprised only interactions (bin_{ij}) in which either *i* or *j* enclose the coordinates of a captured fragment. These interaction matrices were used as input for the reconstruction of 3D models of genomic regions following the modelling protocol described above. The optimal *maxdist* and the *dcutoff* parameters were set at 300 and 200 based on their consensus with the parameters used in the GM12878 pcHi-C dataset. A total of 1,000 steered molecular dynamics trajectories were computed, and for each trajectory the conformations satisfying the majority of the imposed constraints within a radius of 2 σ were retained.

3D chromosome ensemble reconstruction from 'synthetic' sparse dataset

We used a previously published "toy genome" [38] (that is, the ensemble of models accounting for the formation of TAD-like architecture with low structural variability and high noise levels that comprises a total of 626 particles at the highest genomic resolution) to randomly select 10 sets of 22 loci from the toy genome contact map (or synthetic interaction maps). These loci mimic pcHi-C to generate reliable sparse interaction matrices comprising only interactions (bin_{ij}) in which either *i* or *j* have been selected as random captured loci. Each of these sets was then randomly subsampled to generate 'synthetic' capture matrices with 2, 4, 6, 10, 14, and 18 selected captured loci. The obtained 'synthetic' capture matrices (70 in total) were next used as input for the reconstruction of 3D models of genomic regions following the modelling protocol described above. The optimal *maxdist* and the *dcutoff* parameters were set at 500 and 200 nm. Using these parameters, a final ensemble of 100 3D models was reconstructed for each 'synthetic' capture matrices comprising the conformations that best satisfied the imposed restraints in each of the computed 100 steered molecular dynamics trajectories.

Analysis of the ensemble of 3D models

Contact map generation. For each ensemble of 3D models, a contact map was calculated at 5kb resolution to visualize the frequencies of contacts in the ensemble. Two beads were considered

to constitute a contact when their Euclidean distance was below 200 nm cut-off.

Matrix Comparison. The degree of similarity between two matrices was computed by comparing each cell from the matrices, or a subset of them, using the Spearman's rank correlation coefficient (r_s) as implemented in the Python library SciPy [67, 68]:

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(r_{\mathrm{bin}_{x_i}} - r_{\mathrm{bin}_{y_i}})^2}{n(n^2-1)}$$

Where $r_{\mathrm{bin}_{x_i}}$ is the rank of the i^{th} observation in one matrix, $r_{\mathrm{bin}_{y_i}}$ is the rank of the i^{th} observation in the other matrix, and n states for the number of pairs of observations.

Particle-to-particle median distance correlation (ppMdC). For each ensemble of 3D models, we differentiated 3 sets comprising particles enclosing the coordinates of: (i) captured loci (capture), (ii) non-captured loci (other), and (iii) all the loci (all). For each of the pairs of particles in a given set we calculated the particle-to-particle median distance. Then, the degree of similarity between two given sets was computed using the Spearman's rank correlation coefficient between their particle-to-particle median distances. The ppMdC measure varies between −1.0 and 1.0 for comparisons where the particle-to-particle median distances perfectly anti-correlate or correlate, respectively.

Hierarchical clustering of ensembles of 3D models. Multiple ensembles of 3D models were merged in a unique set and the models were structurally superpose using pair-wise rigid-body superposition. Next, the all-vs-all distance root mean square deviation (dRMSD) was calculated and the resulting dRMSD matrix was hierarchically clustered using Ward's sum of squares method [69] as implemented in the Python library SciPy [67].

Cell-specific expression profile. Publicly available [33] expression matrix containing the expression values (log(FPKM)) of each gene in cb-Ery, nCD4, and Mon cell types was downloaded (GeneExpressionMatrix.txt.gz at *https://osf.io/u8tzp/*). The 3 datasets had two or more replicates each (2 cb-Ery, 5 Mac, and 8 nCD4, respectively), thus the average expression value of each gene

from all replicates was used. Then, a cell-specific per-bin cumulative expression profile of the chr11:3,795,000-8,505,000 genomic region at 5kb resolution was obtained assigning the mean expression value of each gene (with log(FPKM)>0) to bins enclosing for the coordinates of its transcription start site (coordinates retrieved from bioMart [70]).

3D enrichment analysis. To study the spatial co-localization of different regulatory elements and the local levels of transcription (based on genome-wide ChIP- and RNA-seq data) around a selected locus (central viewpoint) we implement a *3D enrichment analysis tool* (named 'radial-plot') that allows the comparison of heterogeneous sets of data from multiple data sources. Per each cell type a per-particle binarized chromatin marks profile in the genomic region was generated starting from the ChIP-seq signal of H3K27ac, H3K36me3, H3K4me1, H3K4me3, H3K9me3, and H3K27me3 in cb-Ery, nCD4, and Mon cell types [33]. A particle was considered enclosing for a chromatin mark if a peak was present. Similarly, we also constructed, for each cell type, a per-particle binarized transcription profile starting from the cell-specific expression profile (**Methods**). Then the 3D spatial distribution of the 3D enrichment based on the per-particle binarized profile around the chosen central viewpoint was calculated as follow: (i) starting from the central viewpoint an initial sphere with a radius of 200 nm was constructed; (ii) a series of spherical shells, that occupied a volume equal the initial sphere, were added; (iii) per each model in the ensemble of 3D models a particle of the binarized profile was assigned to a spherical shell based on its relative distance to the central viewpoint; (iv) per each spherical shell we performed Fisher's exact tests for 2 × 2 contingency tables comparing the amount of particles with or without signal in the spherical shell with the outside ones, and the log of the odd ratios was assigned to the shell if the p-value < 0.01. The obtained 3D enrichment was then visualised as a 2D radial plot.

Defining gene communities: co-occurrence of expressed genes. For each ensemble of 3D models, based on their cell-specific expression profile (**Methods**), we defined the set of expressed particles (log(FPKM) > 0). Then, considering this set of particles, an all-vs-all pairwise distances matrix was calculated in each model and hierarchically clustered using Ward's sum of squares method [69]

as implemented in the Python library SciPy [67]. Then the Calinski-Harabasz index [71], as implemented in the Python library Scikit-learn [72], was used to determinate the optimal number of clusters in each dendrogram. Then, for each ensemble, a co-occurrence matrix was generated considering the percentage of models in which a pair of particles belonged to the same cluster. The co-occurrence measure varies between 0 and 100, where 0 indicates absence of co-occurrence and 100 indicates a stable co-occurrence within the ensemble of 3D models. The co-occurrence matrix was next hierarchically clustered using Ward's sum of squares method [69] and communities of co-occurrent active genes were identified using the Calinski-Harabasz index analysis in the dendrogram.

Communities stability within the ensemble of models. To assess the stability of each community within the ensemble we introduced the inter-community co-occurrence score that defines the degree of unstable compositions of a community. It is computed as the mean co-occurrence values between each gene in a community and the rest of the communities.

Distance between communities and within community. To describe the spatial arrangement of each community for a given ensemble of 3D models, we treated each community as a rigid body and calculated its centre of mass (COM) in each 3D model of the ensemble. Per each model the all-vs-all pairwise distances between the COMs of each communities were computed and the mean distance values assigned as the typical distance between communities. Similarly, per each model, we also calculated the distance of each particle in a given community and the COM of its community. The within community distance of a given particle was defined by its mean value in the ensemble of 3D models.

AUTHOR CONTRIBUTIONS

JM-E, IF and MAM-R conceived the study; JM-E, MDS, and IF performed the modeling; DC supported modeling protocol development and implementation; JM-E, IF wrote the manuscript with MDS, DC and MAM-R; IF and MAM-R oversaw the project.

ACKNOWLEDGMENTS

We thank all the current and past members of the Marti-Renom lab for their continuous discussions and support. Dr. Irene Miguel-Escalada for helpful discussions. The 4D genome unit at CRG for data availability. Javierre Lab for providing access to the ChIP-seq peaks for the beta-globin locus in different cell types**.** We acknowledge the ENCODE consortium and the ENCODE production laboratories that generated the datasets used in the manuscript. This study makes use of data generated by the PCHI-C Consortium available in the EGA European Genome-Phenome Archive (National Institute for Health Research of England, UK Medical Research Council (MR/L007150/1) and UK Biotechnology and Biological Research Council (BB/J004480/1)).

FUNDING

This work was partially supported by the European Research Council under the 7th Framework Program FP7/2007-2013 (ERC grant agreement 609989), the European Union's Horizon 2020 research and innovation programme (grant agreement 676556), the Spanish Ministerio de Ciencia, Innovación y Universidades (BFU2013-47736-P and BFU2017-85926-P to M.A.M-R. and IJCI-2015-23352 to I.F), Marató TV3 (201611, to M.A.M-R.). We also knowledge support from "Centro de Excelencia Severo Ochoa 2013-2017", SEV-2012-0208 the Spanish ministry of Science and Innovation to the EMBL partnership and the CERCA Programme/Generalitat de Catalunya to the CRG. We also acknowledge support of the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III, the Generalitat de Catalunya through Departament de Salut and Departament d'Empresa i Coneixement and the Co-financing by the Spanish Ministry of Science and Innovation with funds from the European Regional Development Fund (ERDF) corresponding to the 2014-2020 Smart Growth Operating Program to CNAG.

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289-93. Epub 2009/10/10. doi: 10.1126/science.1181369. PubMed Central PMCID: PMC2858594.
2. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376-80. Epub 2012/04/13. doi: 10.1038/nature11082nature11082 [pii]. PubMed Central PMCID: PMC3356448.
3. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485(7398):381-5. Epub 2012/04/13. doi: 10.1038/nature11049. PubMed Central PMCID: PMC3555144.
4. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012;148(3):458-72. Epub 2012/01/24. doi: 10.1016/j.cell.2012.01.010.
5. Hsieh TS, Cattoglio C, Slobodyanyuk E, Hansen AS, Rando OJ, Tjian R, et al. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. Mol Cell. 2020. doi: 10.1016/j.molcel.2020.03.002.
6. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665-80. doi: 10.1016/j.cell.2014.11.021.
7. Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell. 2017;171(3):557-72 e24. doi: 10.1016/j.cell.2017.09.043. PubMed Central PMCID: PMCPMC5651218.
8. Zheng H, Xie W. The role of 3D genome organization in development and cell differentiation. Nat Rev Mol Cell Biol. 2019. doi: 10.1038/s41580-019-0132-4.

9. Kempfer R, Pombo A. Methods for mapping 3D chromosome architecture. Nat Rev Genet. 2020;21(4):207-26. doi: 10.1038/s41576-019-0195-2.
10. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013;14(6):390-403. doi: 10.1038/nrg3454. PubMed Central PMCID: PMCPMC3874835.
11. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013;502(7469):59-64. doi: 10.1038/nature12593. PubMed Central PMCID: PMCPMC3869051.
12. Ramani V, Deng X, Qiu R, Lee C, Disteche CM, Noble WS, et al. Sci-Hi-C: A single-cell Hi-C method for mapping 3D genome organization in large number of single cells. Methods. 2020;170:61-8. doi: 10.1016/j.ymeth.2019.09.012. PubMed Central PMCID: PMCPMC6949367.
13. Flyamer IM, Gassler J, Imakaev M, Brandao HB, Ulianov SV, Abdennur N, et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature. 2017;544(7648):110-4. doi: 10.1038/nature21711.
14. de Souza N. Genomics. Micro-C maps of genome structure. Nat Methods. 2015;12(9):812. doi: 10.1038/nmeth.3575.
15. Beagrie RA, Scialdone A, Schueler M, Kraemer DC, Chotalia M, Xie SQ, et al. Complex multi-enhancer contacts captured by genome architecture mapping. Nature. 2017;543(7646):519-24. doi: 10.1038/nature21411. PubMed Central PMCID: PMCPMC5366070.
16. Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, et al. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell. 2018;174(3):744-57 e24. doi: 10.1016/j.cell.2018.05.024.
17. van de Werken HJ, de Vree PJ, Splinter E, Holwerda SJ, Klous P, de Wit E, et al. 4C technology: protocols and data analysis. Methods Enzymol. 2012;513:89-112. doi: 10.1016/B978-0-12-391938-0.00004-5.
18. Allahyar A, Vermeulen C, Bouwman BAM, Krijger PHL, Verstegen M, Geeven G, et al. Enhancer hubs and loop collisions identified from single-allele topologies. Nat

Genet. 2018;50(8):1151-60. doi: 10.1038/s41588-018-0161-5.
19. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods. 2016;13(11):919-22. doi: 10.1038/nmeth.3999.
20. Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre BM, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. 2015;25(4):582-97. Epub 2015/03/11. doi: 10.1101/gr.185272.114. PubMed PMID: 25752748; PubMed Central PMCID: PMC4381529.
21. Bendandi A, Dante S, Zia SR, Diaspro A, Rocchia W. Chromatin Compaction Multiscale Modeling: A Complex Synergy Between Theory, Simulation, and Experiment. Front Mol Biosci. 2020;7:15. doi: 10.3389/fmolb.2020.00015. PubMed Central PMCID: PMCPMC7051991.
22. Oluwadare O, Highsmith M, Cheng J. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. Biol Proced Online. 2019;21:7. doi: 10.1186/s12575-019-0094-0. PubMed Central PMCID: PMCPMC6482566.
23. Serra F, Di Stefano M, Spill YG, Cuartero Y, Goodstadt M, Bau D, et al. Restraint-based three-dimensional modeling of genomes and genomic domains. FEBS Lett. 2015;589(20 Pt A):2987-95. doi: 10.1016/j.febslet.2015.05.012.
24. Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, et al. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol. 2011;18(1):107-14. Epub 2010/12/07. doi: 10.1038/nsmb.1936 [pii]. PubMed Central PMCID: PMC3056208.
25. Tjong H, Li W, Kalhor R, Dai C, Hao S, Gong K, et al. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proc Natl Acad Sci U S A. 2016;113(12):E1663-72. doi: 10.1073/pnas.1512577113. PubMed Central PMCID: PMCPMC4812752.
26. Hua N, Tjong H, Shin H, Gong K, Zhou XJ, Alber F. Producing genome structure populations with the dynamic

and automated PGS software. Nat Protoc. 2018;13(5):915-26. doi: 10.1038/nprot.2018.008. PubMed Central PMCID: PMCPMC6043163.

27. Serra F, Bau D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput Biol. 2017;13(7):e1005665. doi: 10.1371/journal.pcbi.1005665. PubMed Central PMCID: PMCPMC5540598.
28. Irastorza-Azcarate I, Acemel RD, Tena JJ, Maeso I, Gomez-Skarmeta JL, Devos DP. 4Cin: A computational pipeline for 3D genome modeling and virtual Hi-C analyses from 4C data. PLoS Comput Biol. 2018;14(3):e1006030. doi: 10.1371/journal.pcbi.1006030. PubMed Central PMCID: PMCPMC5862518.
29. Di Stefano M, Stadhouders R, Farabella I, Castillo D, Serra F, Graf T, et al. Transcriptional activation during cell reprogramming correlates with the formation of 3D open chromatin hubs. Nat Commun. 2020;11(1):2564. doi: 10.1038/s41467-020-16396-1. PubMed Central PMCID: PMCPMC7244774.
30. Paulsen J, Gramstad O, Collas P. Manifold Based Optimization for Single-Cell 3D Genome Reconstruction. PLoS Comput Biol. 2015;11(8):e1004396. doi: 10.1371/journal.pcbi.1004396. PubMed Central PMCID: PMCPMC4532452.
31. Stevens TJ, Lando D, Basu S, Atkinson LP, Cao Y, Lee SF, et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature. 2017;544(7648):59-64. doi: 10.1038/nature21429. PubMed Central PMCID: PMCPMC5385134.
32. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015;47(6):598-606. doi: 10.1038/ng.3286.
33. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016;167(5):1369-84 e19. doi:

10.1016/j.cell.2016.09.037. PubMed Central PMCID: PMCPMC5123897.
34. Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet. 2019;51(10):1442-9. doi: 10.1038/s41588-019-0494-8. PubMed Central PMCID: PMCPMC6778519.
35. Miguel-Escalada I, Bonas-Guarch S, Cebola I, Ponsa-Cobas J, Mendieta-Esteban J, Atla G, et al. Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. Nat Genet. 2019;51(7):1137-48. doi: 10.1038/s41588-019-0457-0. PubMed Central PMCID: PMCPMC6640048.
36. Russel D, Lasker K, Webb B, Velazquez-Muriel J, Tjioe E, Schneidman-Duhovny D, et al. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol. 2012;10(1):e1001244. Epub 2012/01/25. doi: 10.1371/journal.pbio.1001244 [pii]. PubMed Central PMCID: PMC3260315.
37. Vidal E, le Dily F, Quilez J, Stadhouders R, Cuartero Y, Graf T, et al. OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. Nucleic Acids Res. 2018;46(8):e49. doi: 10.1093/nar/gky064. PubMed Central PMCID: PMCPMC5934634.
38. Trussart M, Serra F, Bau D, Junier I, Serrano L, Marti-Renom MA. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. Nucleic Acids Res. 2015;43(7):3465-77. doi: 10.1093/nar/gkv221. PubMed Central PMCID: PMCPMC4402535.
39. Palstra RJ, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, de Laat W. The beta-globin nuclear compartment in development and erythroid differentiation. Nat Genet. 2003;35(2):190-4.
40. Levings PP, Bungert J. The human beta-globin locus control region. Eur J Biochem. 2002;269(6):1589-99. doi: 10.1046/j.1432-1327.2002.02797.x.
41. Liu X, Zhang Y, Chen Y, Li M, Zhou F, Li K, et al. In Situ Capture of Chromatin Interactions by Biotinylated dCas9. Cell. 2017;170(5):1028-43 e19. doi:

10.1016/j.cell.2017.08.003. PubMed Central PMCID: PMCPMC6857456.
42. Fraser P, Pruzina S, Antoniou M, Grosveld F. Each hypersensitive site of the human beta-globin locus control region confers a different developmental pattern of expression on the globin genes. Genes Dev. 1993;7(1):106-13. doi: 10.1101/gad.7.1.106.
43. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. Nature. 2007;447(7143):413-7. Epub 2007/05/25. doi: nature05916 [pii] 10.1038/nature05916.
44. Jackson DA, Hassan AB, Errington RJ, Cook PR. Visualization of focal sites of transcription within human nuclei. EMBO J. 1993;12(3):1059-65. Epub 1993/03/01.
45. Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet. 2004;36(10):1065-71.
46. Osborne CS, Chakalova L, Mitchell JA, Horton A, Wood AL, Bolland DJ, et al. Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. PLoS Biol. 2007;5(8):e192. Epub 2007/07/12. doi: 06-PLBI-RA-2184 [pii] 10.1371/journal.pbio.0050192.
47. Ben Zouari Y, Molitor AM, Sikorska N, Pancaldi V, Sexton T. ChiCMaxima: a robust and simple pipeline for detection and visualization of chromatin looping in Capture Hi-C. Genome Biol. 2019;20(1):102. doi: 10.1186/s13059-019-1706-3. PubMed Central PMCID: PMCPMC6532271.
48. Cairns J, Freire-Pritchett P, Wingett SW, Varnai C, Dimond A, Plagnol V, et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. 2016;17(1):127. doi: 10.1186/s13059-016-0992-2. PubMed Central PMCID: PMCPMC4908757.
49. Cairns J, Orchard WR, Malysheva V, Spivakov M. Chicdiff: a computational pipeline for detecting differential chromosomal interactions in Capture Hi-C data. Bioinformatics. 2019;35(22):4764-6. doi: 10.1093/bioinformatics/btz450. PubMed Central PMCID: PMCPMC6853696.
50. Mifsud B, Martincorena I, Darbo E, Sugar R, Schoenfelder S, Fraser P, et al. GOTHiC, a probabilistic model to resolve

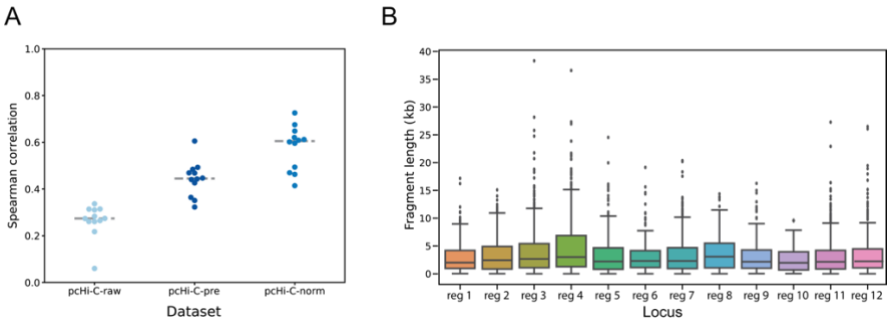
complex biases and to identify real interactions in Hi-C data. PLoS One. 2017;12(4):e0174744. doi: 10.1371/journal.pone.0174744. PubMed Central PMCID: PMCPMC5381888.

51. Anil A, Spalinskas R, Akerborg O, Sahlen P. HiCapTools: a software suite for probe design and proximity detection for targeted chromosome conformation capture applications. Bioinformatics. 2018;34(4):675-7. doi: 10.1093/bioinformatics/btx625. PubMed Central PMCID: PMCPMC6368139.
52. Brown JM, Leach J, Reittie JE, Atzberger A, Lee-Prudhoe J, Wood WG, et al. Coregulated human globin genes are frequently in spatial proximity when active. J Cell Biol. 2006;172(2):177-87. doi: 10.1083/jcb.200507073. PubMed Central PMCID: PMCPMC2063548.
53. Schubeler D, Francastel C, Cimbora DM, Reik A, Martin DI, Groudine M. Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human beta-globin locus. Genes Dev. 2000;14(8):940-50. PubMed Central PMCID: PMCPMC316536.
54. Huang P, Keller CA, Giardine B, Grevet JD, Davies JOJ, Hughes JR, et al. Comparative analysis of three-dimensional chromosomal architecture identifies a novel fetal hemoglobin regulatory element. Genes Dev. 2017;31(16):1704-13. doi: 10.1101/gad.303461.117. PubMed Central PMCID: PMCPMC5647940.
55. Sanyal A, Bau D, Marti-Renom MA, Dekker J. Chromatin globules: a common motif of higher order chromosome structure? Curr Opin Cell Biol. 2011;23(3):325-31. doi: 10.1016/j.ceb.2011.03.009. PubMed Central PMCID: PMCPMC3109114.
56. Sutherland H, Bickmore WA. Transcription factories: gene expression in unions? Nat Rev Genet. 2009;10(7):457-66. Epub 2009/06/10. doi: nrg2592 [pii] 10.1038/nrg2592.
57. Iborra FJ, Pombo A, Jackson DA, Cook PR. Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. J Cell Sci. 1996;109 (Pt 6):1427-36.
58. Gurumurthy A, Shen Y, Gunn EM, Bungert J. Phase Separation and Transcription Regulation: Are Super-

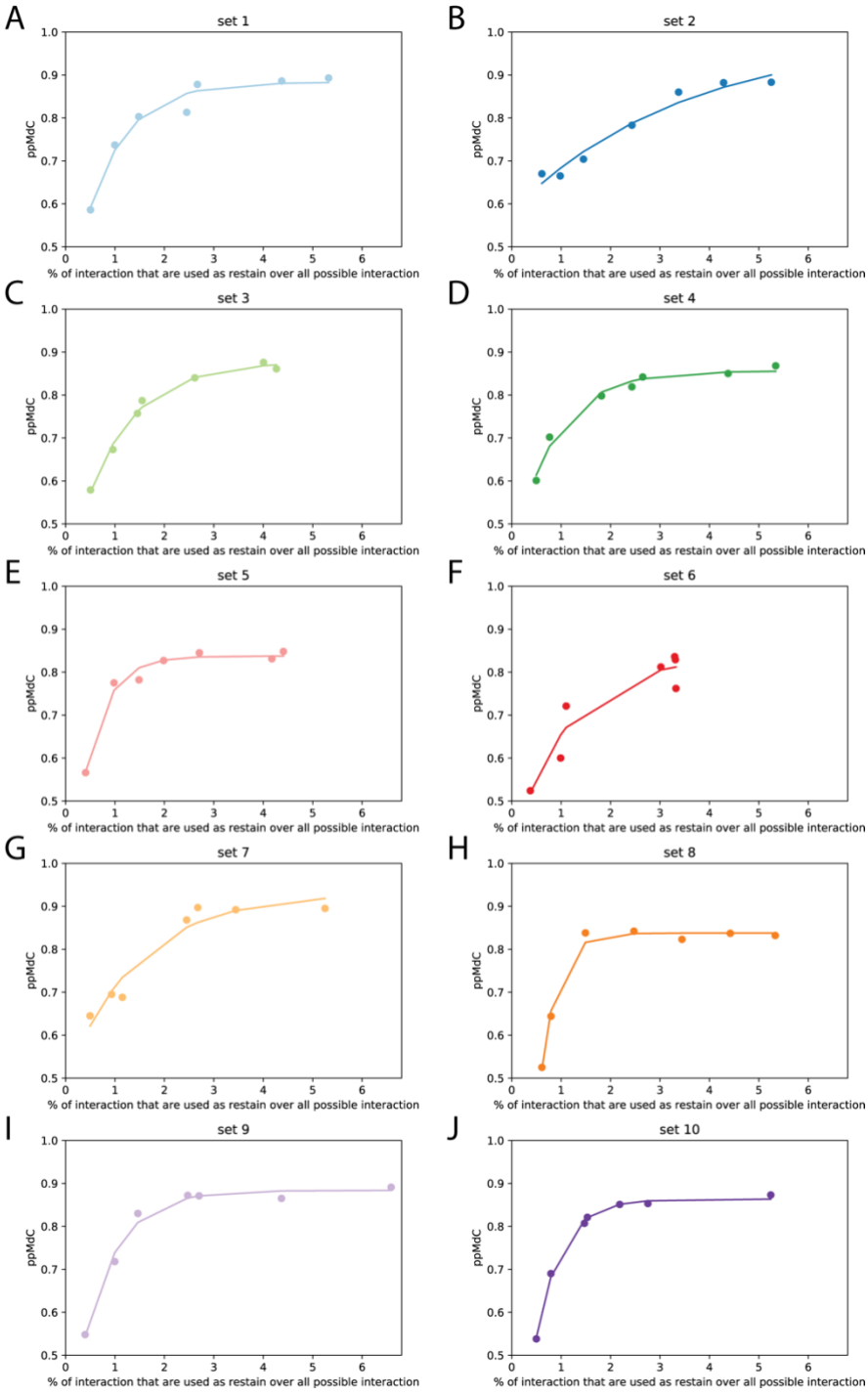
Enhancers and Locus Control Regions Primary Sites of Transcription Complex Assembly? Bioessays. 2019;41(1):e1800164. doi: 10.1002/bies.201800164. PubMed Central PMCID: PMCPMC6484441.
59. Boija A, Klein IA, Sabari BR, Dall'Agnese A, Coffey EL, Zamudio AV, et al. Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. Cell. 2018;175(7):1842-55 e16. doi: 10.1016/j.cell.2018.10.042. PubMed Central PMCID: PMCPMC6295254.
60. Cho WK, Spille JH, Hecht M, Lee C, Li C, Grube V, et al. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science. 2018;361(6400):412-5. doi: 10.1126/science.aar4199. PubMed Central PMCID: PMCPMC6543815.
61. Yang T, Zhang F, Yardimci GG, Song F, Hardison RC, Noble WS, et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res. 2017;27(11):1939-49. doi: 10.1101/gr.220640.117. PubMed Central PMCID: PMCPMC5668950.
62. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012;9(10):999-1003. Epub 2012/09/04. doi: 10.1038/nmeth.2148. PubMed Central PMCID: PMC3816492.
63. Di Stefano M, Rosa A, Belcastro V, di Bernardo D, Micheletti C. Colocalization of coregulated genes: a steered molecular dynamics study of human chromosome 19. PLoS Comput Biol. 2013;9(3):e1003019. doi: 10.1371/journal.pcbi.1003019. PubMed Central PMCID: PMCPMC3610629.
64. Kremer K, Grest GS. Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. The Journal of Chemical Physics. 1990;92(8):5057-86. doi: 10.1063/1.458541.
65. Rosa A, Everaers R. Structure and dynamics of interphase chromosomes. PLoS Comput Biol. 2008;4(8):e1000153. doi: 10.1371/journal.pcbi.1000153. PubMed Central PMCID: PMCPMC2515109.

66. Polak E, Ribiere G. Note sur la convergence de méthodes de directions conjuguées. ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique. 1969;3:35-43.
67. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261-72. doi: 10.1038/s41592-019-0686-2. PubMed Central PMCID: PMCPMC7056644.
68. Zwillinger D, Kokoska S. CRC standard probability and statistics tables and formulae. Boca Raton: Chapman & Hall/CRC; 2000.
69. Ward JH. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association. 1963;58(301):236-44. doi: 10.1080/01621459.1963.10500845.
70. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015;43(W1):W589-98. doi: 10.1093/nar/gkv350. PubMed Central PMCID: PMCPMC4489294.
71. Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics. 1974;3(1):1-27. doi: 10.1080/03610927408827101.
72. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(null):2825–30.
73. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605-12.

SUPPLEMENTARY FIGURES AND TABLES

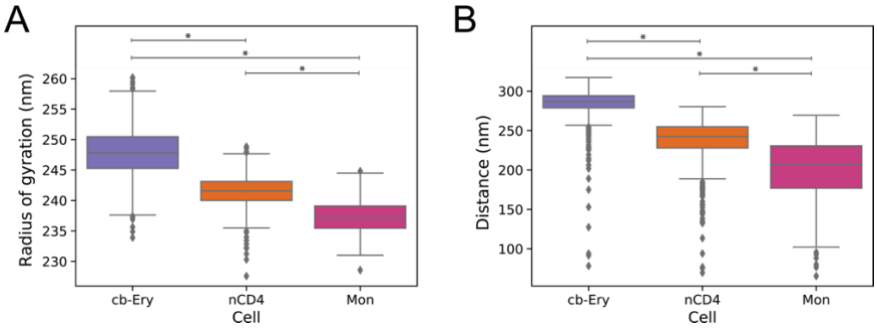


Supplementary Figure 1. Integrative modelling procedure: assessing Print normalisation procedure and defining the model representation. (A) Assessing Print multi-step procedure. Element-wise Spearman correlation between each stage of Print normalisation (pcHi-C-raw in light blue; pcHi-C-pre in dark blue; and pcHi-C-norm in medium blue) and the Hi-C interaction matrix. The grey dashed line indicates the median correlation in the entire benchmark dataset at each stage. **(B)** Distribution of the sizes (in kb) of the restriction fragments in each of the regions comprised in the benchmark dataset.

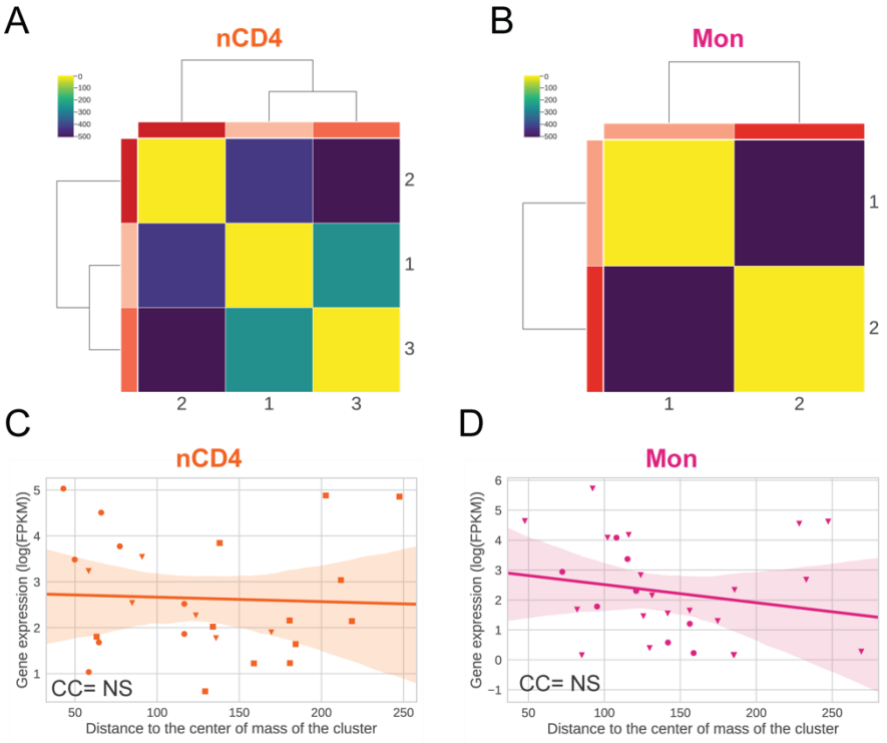


Supplementary Figure 2. Comparison of each 'synthetic' capture sets and the toy genome. (A-J) Relationship between the ppMdC and the percentage of cells in

the matrix used as restraints in each set. The dots in the plot represent the degree of sparseness in each subset (2, 4, 6, 10, 14, 18, and 22 captures) and the coloured line indicates the fitted exponential function. The colour code used is the same as in Figure 2C.



Supplementary Figure 3. Cell-type-specific 3D features in the model ensembles. (A) Cell-type specific distribution of the radius of gyration of the models in the ensemble. Box boundaries represent 1st and 3rd quartiles, middle line represents median, and whiskers extend to 1.5 times the interquartile range (two-samples Kolmogorov-Smirnov test, asterisk indicate $p < 9.1e^{-163}$). **(B)** Cell-type specific distance distribution of the centres of mass of the particles containing the β-globin genes (HBB, HBD, HBG1, HBG2, and HBE1) from the centre of the model as calculated in each model of the ensemble. Box boundaries represent 1st and 3rd quartiles, middle line represents median, and whiskers extend to 1.5 times the interquartile range (two-samples Kolmogorov-Smirnov test, asterisk indicate $p < 3.46e^{-101}$).



Supplementary Figure 4. Hierarchical 3D organisation of expressed entities in nCD4 and Mon. (A-B) Hierarchical clustering of the distances of mass (**Methods**) between the communities defined in nCD4 **(A)** and Mon **(B)**. Distance value is coloured in the matrix from dark blue (low) to bright yellow (high) and the average expression in log(FPKM) per community is coloured by ranking from lowest (lightest) to highest (darkest) in 3 **(A)** and 2 **(B)** different shades of red. **(C-D)** Relationship between gene expression in log(FPKM) and the median distance of the gene particles to the centre of mass of its own community in nCD4 **(C)** and Mon **(D)** ensemble of models (**Methods**). Orange **(C)** and pink **(D)** line denote the linear regression fit, the shading around the regression line represents the confidence interval, each community is represented with different symbols (circle community 1; inverse triangle community 2; square community 3; and ex community 4) ; NS stand for not significant.

Supplementary Table 1. Benchmark datasets

Locus	Chromosome	Start	End
Region 1	chr7	137,515,000	138,120,000
Region 2	chr8	132,755,000	133,560,000
Region 3	chr20	50,745,000	52,515,000
Region 4	chr3	49,325,000	51,095,000
Region 5	chr3	63,110,000	64,715,000
Region 6	chr8	119,785,000	120,190,000
Region 7	chr17	68,500,000	70,005,000
Region 8	chr2	10,705,000	11,210,000
Region 9	chr10	88,890,000	89,495,000
Region 10	chr1	169,590,000	169,745,000
Region 11	chr21	26,625,000	28,930,000
Region 12	chr13	84,575,000	86,180,000

Description: **Locus**, the name of the region modelled starting from Hi-C, pcHi-C, and pcHi-Cvirt datasets; **Chromosome**, the chromosome where the region is located; **Start** and **End**, represent the genomic coordinates (GRCh38 assembly).

Supplementary Table 2. Gene communities expression statistics

Cell	Community	nGenes	MeanExp	σ
cb-Ery	1	8	2.48	0.81
	2	11	4.57	3.28
	3	3	1.57	1.99
	4	6	2.78	0.78
NCD4	1	8	2.48	0.81
	2	6	4.57	3.28
	3	12	1.57	1.99
Mon	1	8	2.48	0.81
	2	20	4.57	3.28

Description: **Cell**, source cell type data used for the modelling approach; **Community**, the number assigned to each community; **nGenes** , number of active genes composing each community; **MeanExp**, mean expression of the genes within the community; **σ**, standard deviation of the mean expression of genes within the community.

Supplementary Table 3. MMP scores of the 12 modelled Hi-C interaction matrices

Dataset	Locus	MMP score
HiC data (GSE63525)	Region 1	0.80
	Region 2	0.78
	Region 3	0.76
	Region 4	0.78
	Region 5	0.78
	Region 6	0.81
	Region 7	0.78
	Region 8	0.81
	Region 9	0.82
	Region 10	0.84
	Region 11	0.72
	Region 12	0.75

Description: **Dataset**, experimental dataset used to reconstruct the 12 ensembles of models; **Locus**, the name of the region modelled starting from the Hi-C dataset; **MMP score**, Value of the MMP score of the interaction matrix of each of the locus. It predicts the reliability of the 3D models based on the interaction matrix size, the contribution of significant eigenvectors in the matrix, and the skewness and kurtosis of the z-scores distribution of the matrix [1, 2].

1. Trussart M, Serra F, Bau D, Junier I, Serrano L, Marti-Renom MA. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. Nucleic Acids Res. 2015;43(7):3465-77. doi: 10.1093/nar/gkv221. PubMed Central PMCID: PMCPMC4402535.
2. Serra F, Bau D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput Biol. 2017;13(7):e1005665. doi: 10.1371/journal.pcbi.1005665. PubMed Central PMCID: PMCPMC5540598.

CHAPTER 2

Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes

Irene Miguel-Escalada, Silvia Bonàs-Guarch, Inês Cebola, Joan Ponsa-Cobas, Julen Mendieta-Esteban, Goutham Atla, Biola M. Javierre, Delphine M. Y. Rolando, Irene Farabella, Claire C. Morgan, Javier García-Hurtado, Anthony eucher, Ignasi Morán, Lorenzo Pasquali, Mireia Ramos-Rodríguez, Emil V. R. Appel, Allan Linneberg, Anette P. Gjesing, Daniel R. Witte, Oluf Pedersen, Niels Grarup, Philippe avassard, David Torrents, Josep M. Mercader, Lorenzo iemonti, Thierry Berney, Eelco J. P. de Koning, Julie Kerr-Conte, François Pattou, Iryna O. Fedko, Leif Groop, Inga Prokopenko 28,29, Torben Hansen 10, Marc A. Marti-Renom, Peter Fraser and Jorge Ferrer. **Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes.** Nature Genetics. 51, 1137–1148 (2019).

Human pancreatic islet 3D chromatin architecture provides insights into the genetics of type 2 diabetes

Authors:
Irene Miguel-Escalada[1-4,32], Silvia Bonàs-Guarch[1-4,32], Inês Cebola[1,32], Joan Ponsa-Cobas[1], Julen Mendieta-Esteban[5], Goutham Atla[1-4], Biola M. Javierre[6,7], Delphine M.Y. Rolando[1], Irene Farabella[5], Claire C. Morgan[1,4], Javier García-Hurtado[2-4], Anthony Beucher[1], Ignasi Morán[1,16], Lorenzo Pasquali[2,7,8], Mireia Ramos-Rodríguez[8], Emil V.R. Appel[9], Allan Linneberg[10,11], Anette P. Gjesing[9], Daniel R. Witte[12,13], Oluf Pedersen[9], Niels Grarup[9], Philippe Ravassard[14], David Torrents[15,16], Josep M. Mercader[16-18], Lorenzo Piemonti[19,20] , Thierry Berney[21], Eelco J.P. de Koning[22,23], Julie Kerr-Conte[24], François Pattou[24], Iryna O. Fedko[25,26], Leif Groop[27], Inga Prokopenko[28,29], Torben Hansen[9], Marc A. Marti-Renom[5,15,30], Peter Fraser[6,31], Jorge Ferrer[1,2,4,*]

Affiliations
1- Section of Epigenomics and Disease, Department of Medicine, and National Institute for Health Research (NIHR) Imperial Biomedical Research Centre, Imperial College London, London W12 0NN, UK
2- CIBER de Diabetes y Enfermedades Metabólicas Asociadas, Spain
3- Genomic Programming of Beta-cells Laboratory, Institut d'Investigacions August Pi i Sunyer (IDIBAPS), 08036 Barcelona, Spain
4- Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
5- CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri Reixac 4, Barcelona 08028, Spain
6- Nuclear Dynamics Programme, The Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, UK
7- Josep Carreras Leukaemia Research Institute, Campus ICO-Germans Trias i Pujol, Ctra de Can Ruti, Camí de les Escoles s/n, Badalona, 08916, Spain

8- Endocrine Regulatory Genomics Lab, Germans Trias i Pujol University Hospital and Research Institute, 08916 Badalona, Spain
9- Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
10- Center for Clinical Research and Disease Prevention, Bispebjerg and Frederiksberg Hospital, The Capital Region, Copenhagen, Denmark.
11- Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
12- Department of Public Health, Aarhus University, Aarhus, Denmark.
13- Danish Diabetes Academy, Odense, Denmark
14- Université Sorbonne, UPMC Univ Paris 06, Inserm, CNRS, Institut du cerveau et de la moelle (ICM) – Hôpital Pitié-Salpêtrière, Boulevard de l'Hôpital, Paris F-75013, France
15- ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain
16- Barcelona Supercomputing Center (BSC), Joint BSC-CRG-IRB Research Program in Computational Biology, 08034 Barcelona, Spain
17- Programs in Metabolism and Medical & Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
18- Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
19- Diabetes Research Institute (SR-DRI), IRCCS San Raffaele Scientific Institute, Via Olgettina 60, 20132 Milan, Italy
20- Vita-Salute San Raffaele University, Milan
21- Cell Isolation and Transplantation Center, University of Geneva, 1211 Geneva 4, Switzerland
22- Department of Medicine, Leiden University Medical Center, Box 9600, 2300 RC Leiden, Netherlands
23- Hubrecht Institute/KNAW, 85164 3508 AD Utrecht, the Netherlands
24- European Genomic Institute for Diabetes, Inserm UMR 1190, Lille 59800, France
25- Department of Biological Psychology, Vrije Universiteit Amsterdam, The Netherlands
26- Amsterdam Public Health research institute, The Netherlands

27- Genomics, Diabetes and Endocrinology, Department of Clinical Sciences, Clinical Research Centre, Lund University, Malmö, Sweden
28- Section of Genomics of Common Disease, Department of Medicine, Imperial College London, London W12 0NN, UK
29- Department of Clinical and Experimental Medicine, University of Surrey, Guildford, GU2 7XH, UK
30- Universitat Pompeu Fabra (UPF), Barcelona, Spain
31- Department of Biological Science, Florida State University, Tallahassee, Florida 32303, USA
32-These authors contributed equally, Irene Miguel-Escalada, Silvia Bonàs-Guarch, Inês Cebola
* e-mail: *j.ferrer@imperial.ac.uk*

ABSTRACT

Genetic studies promise to provide insight into the molecular mechanisms underlying type 2 diabetes (T2D). Variants associated with T2D are often located in tissue-specific enhancer clusters or super-enhancers. So far, such domains have been defined through clustering of enhancers in linear genome maps rather than in 3D space. Furthermore, their target genes are often unknown. We have created promoter capture Hi-C maps in human pancreatic islets. This linked diabetes-associated enhancers to their target genes, often located hundreds of kilobases away. It also revealed > 1,300 groups of islet enhancers, super-enhancers and active promoters that form 3D hubs, some of which show coordinated glucose-dependent activity. We demonstrate that genetic variation in hubs impacts insulin secretion heritability, and show that hub annotations can be used for polygenic scores that predict T2D risk driven by islet regulatory variants. Human islet 3D chromatin architecture, therefore, provides a framework for interpretation of T2D GWAS signals.

INTRODUCTION

Type 2 diabetes (T2D) affects more than 400 million people worldwide [1], and is a classic example of a polygenic disease in which the genetic susceptibility is largely driven by noncoding variants[2,3]. T2D susceptibility variants are enriched in active islet enhancers that cluster in linear genome maps – variably defined as

super-enhancers, COREs, enhancer clusters, or stretch enhancers [4-7]. Enhancer clusters from other tissues or cell types are similarly enriched in risk variants for various common diseases[5,7-11]. So far, however, genome-wide maps of enhancer clusters have been largely defined with unidimensional epigenomic maps, which do not necessarily reflect the capacity of enhancers to cluster in three-dimensional (3D) space, as shown for well characterized loci such as *Hbb* (β-globin) and *Hoxd*[12,13]. Linear maps also do not reveal the target genes of enhancers, which are often separated by hundreds of thousands of base pairs. Therefore, there is a need to obtain accurate representations or enhancer domains, and to connect them to the target genes that underpin disease mechanisms.

Here, we used promoter capture Hi-C (pcHi-C) [14] to generate a genome-scale map of interactions between gene promoters and their regulatory elements in human pancreatic islets. This uncovered ~1,300 hubs of islet enhancers that cluster in 3D space. We show that islet enhancer hubs are connected with key islet gene promoters, and exhibit properties of regulatory domains. We use genome/epigenome editing to demonstrate the functional connectivity of hubs, and validate functional interactions between enhancers bearing T2D risk variants and their target genes. Finally, we show that islet hubs are not only enriched for T2D association signals, but can be used to partition polygenic scores to identify T2D genetic susceptibility driven by pancreatic islet regulatory variation.

RESULTS

The promoter interactome of human islets.

To create a genome-wide, high-resolution map of long-range interactions between gene promoters and distant regulatory elements in human pancreatic islets, we prepared Hi-C libraries from four human islet samples, and then performed hybridization capture of 31,253 promoter-containing HindIII fragment baits and their ligated DNA fragments. These were then sequenced and processed with the CHiCAGO algorithm to define 175,784 high-confidence interactions (CHiCAGO score > 5) between annotated promoters and distal genomic regions promoter-interacting DNA fragments [14,15] (Fig. 1a,b and Supplementary Fig. 1). These high-confidence interactions were called with pooled samples, but for

89% of interactions all individual samples showed CHiCAGO scores above the 95% confidence interval of random distance-matched regions (Supplementary Fig. 1d-g). We also validated pcHi-C landscapes by 4C-seq analysis in the EndoC-βH1 human β cell line in two selected loci (Supplementary Fig. 1h,i).

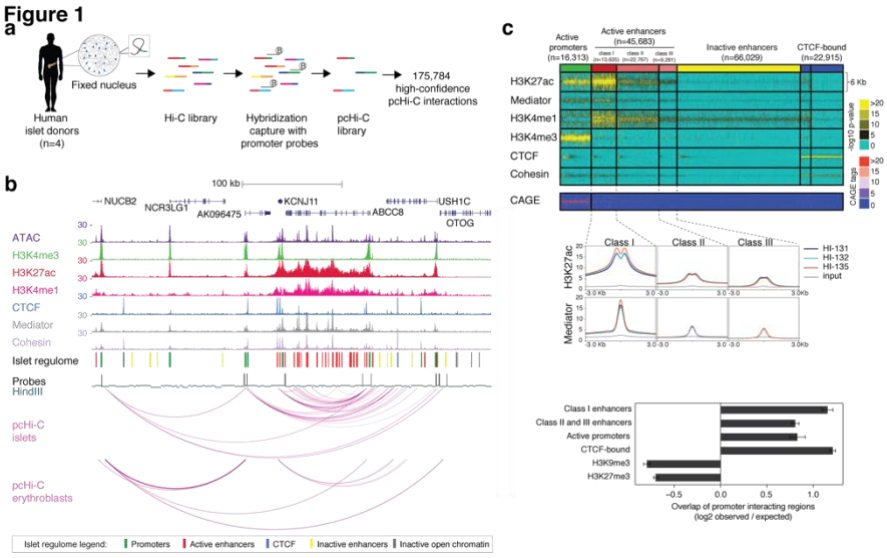


Figure 1. The promoter interactome of human pancreatic islets. a, Overview of promoter-capture Hi-C (pcHi-C) in human islets. b, Integrative map of the *KCNJ11-ABCC8* locus, showing human islet ATAC-seq and ChIP-seq, HindIII bait fragments, and arcs representing high-confidence pcHi-C interactions in human islets and erythroblasts. c, High-resolution annotations of islet open chromatin. ATAC-seq data from 13 islet samples were used to define consistent open chromatin regions, which were classified with k-medians clustering based on epigenomic features. Mediator and H3K27ac binding patterns allowed sub-classification of active enhancer classes I-III. Post-hoc analysis of islet CAGE tags confirmed that transcription start sites are highly enriched in promoters and weakly in class I enhancers. These islet *regulome* annotations are hereafter Supplementary Data Set 1. d, Average H3K27ac and Mediator signal centered on open chromatin regions for active enhancer subtypes in three human islet (HI) samples and input DNA. e, Overlap of promoter-interacting regions with epigenomic features, expressed as average $\log_2$ ratios (and 95% confidence intervals) over the overlaps obtained with 100 sets of distance-matched fragments. Error bars show s.d. across control sets.

To define the chromatin landscape of interacting regions, we refined existing human islet epigenome annotations by generating human islet ATAC-seq maps and 30 new ChIP-seq datasets (Fig. 1b-d, Supplementary Table 1). This enabled a subclassification of active

enhancers according to Mediator, cohesin, and H3K27ac occupancy patterns (Fig. 1b-d, Supplementary Data Set 1). Expectedly, promoter-interacting genomic regions were enriched in active enhancers, promoters, and CTCF-bound regions (Fig. 1e, Supplementary Fig. 2a-c). pcHi-C interactions observed in pcHi-C maps from distant cell types were enriched in CTCF binding sites and active promoters, whereas islet-selective interacting regions were enriched in active enhancers (particularly those with strongest Mediator occupancy, which we term class I enhancers) and were connected with genes showing islet-specific expression (Supplementary Fig. 2d-f). This genome-scale map of the human pancreatic islet promoter interactome is accessible for visualization along with pcHi-C maps of other human tissues (*www.chicp.org*) [16], or as virtual 4C representations of all genes along with islet regulatory annotations (isletregulome.org) [17].

Identification of target genes for islet enhancers.
Long-range chromatin interactions are largely constrained within topologically associating domains (TADs), which typically span hundreds of kilobases and are often invariant across tissues (Supplementary Fig. 3a-e) [18,19]. TADs, however, define broad genomic intervals that do not necessarily inform on the specific interactions that take place in each tissue between individual *cis*-regulatory elements and their target genes. Human islet pcHi-C maps identified high-confidence pcHi-C interactions (CHiCAGO score > 5) between gene promoters and 18,031 different islet enhancers (Fig. 2a). Remarkably, 42.2% of enhancers that showed interactions with gene promoters had high-confidence interactions with more than one gene, thereby illustrating an unexpected complexity of islet enhancer-promoter interactions (Supplementary Fig. 3f).

We used pcHi-C maps to further expand the number of enhancers that could be assigned to target genes. We reasoned that interactions between enhancers and their target genes can be missed due the stringency of detection thresholds, the strong bias of Hi-C methods against proximal interactions, or their dependence on specific environmental conditions. To impute additional enhancer-promoter assignments, we considered promoter-associated three-dimensional spaces (PATs). A PAT was defined as the space containing all pcHi-C interactions that stem from a promoter bait (Supplementary

Fig. 3g,h). We observed that PATs that had one high-confidence enhancer-promoter interaction were more likely to show other enhancer-promoter interactions, and exhibited chromatin features that distinguished them from other PATs (Supplementary Fig. 3 i-k). This prompted us to leverage PAT features to impute plausible target promoter(s) of an additional 18,633 islet enhancers that did not show high-confidence interactions (Fig. 2a; see Supplementary Fig. 3l and Methods for a detailed description of the imputation pipeline). Imputed promoter-enhancer pairs showed higher CHiCAGO scores than distance-matched regions (Kruskall-Wallis $P < 10^{-16}$), suggesting that many imputed assignments represent physical interactions that do not reach our stringent significance thresholds (Supplementary Fig. 3m). In total, we used high-confidence interactions and imputations to assign 36,664 human islet active enhancers (80% of all enhancers) to at least one target gene (Fig. 2a, Supplementary Data Set 2).

We validated these enhancer-to-gene assignments with complementary approaches. First, we calculated normalized H3K27ac signals in assigned enhancer-promoter pairs across human tissues and human islet samples, and found that assigned pairs had distinctly higher correlation values than enhancers paired with distance-matched promoters from the same TAD or an overlapping PAT (Fig. 2b). Importantly, this was true for both high-confidence and imputed assignments (Fig. 2b). Islet-selective expression was expectedly enriched in enhancer-assigned genes but not in unassigned genes from the same TAD (Supplementary Fig. 3n). Furthermore, we determined 1,091 eQTL-genes (eGenes) from 183 human islet samples (Supplementary Table 2), and found that eQTLs were enriched in enhancer-to-gene assignments determined through either high-confident interactions or imputations, compared with distance-matched regions (odds ratio 3.18 and 4.36; $P = 3.05 \times 10^{-9}$ and 9.01×10^{-23}, respectively) (Fig. 2c).

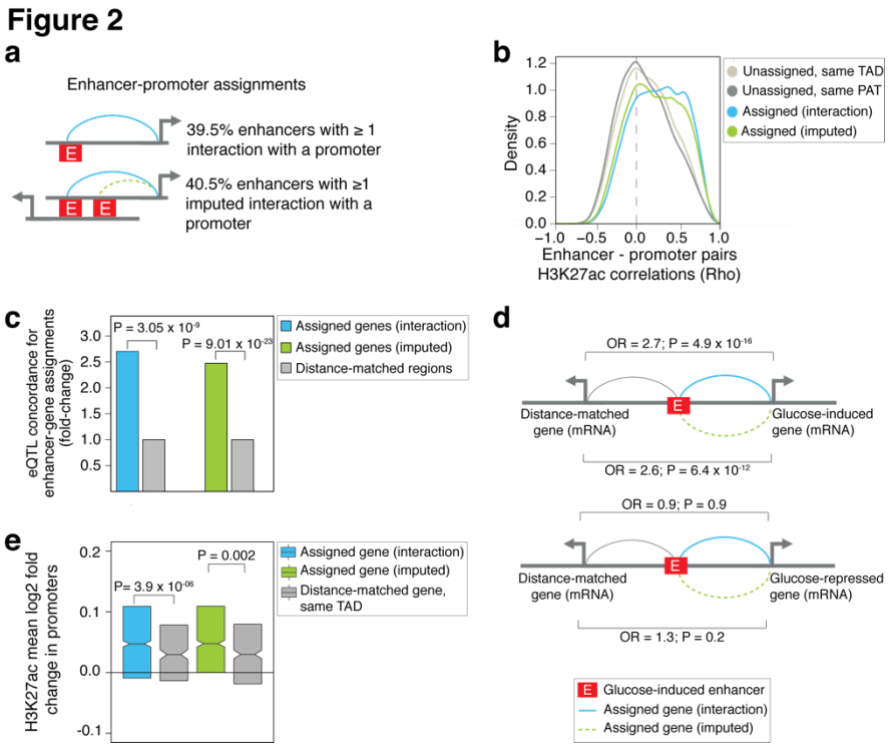


Figure 2. Identification of target genes of islet enhancers. a, We assigned target genes to 39.5% of all 45,683 active enhancers through high-confidence interactions. PAT features allowed imputing the assignment of promoters to another 40% of all active enhancers (see Supplementary Fig. 3l,m for further details and evidence that imputed assigments are enriched in sub-threshold interactions). b, Functional correlation of enhancer-gene pairs assigned through high-confidence interactions (n = 18,637 pairs) or imputations (n = 28,695 pairs). Spearman's Rho values for normalized H3K27ac signal in enhancer-promoter pairs across 14 human islet samples and 51 Roadmap Epigenomics tissues. Control enhancer-gene pairs were enhancers that overlapped a PAT in linear maps but were not assigned to the PAT promoter (n = 9,770 pairs), or other unassigned gene-enhancer pairs from the same TAD (n = 20,186 pairs). c, Concordance of enhancer eQTL-eGene pairs and enhancers-gene pairs assigned through high-confidence interactions (n = 351 pairs) or imputations (n = 293 pairs), relative to distance-matched control regions (n = 579 and 593 pairs, respectively), shown as a fold-change. *P* values were derived from one-sided Fisher's exact test. d, Genes assigned to glucose-induced enhancers show concordant glucose-induced expression. Top: glucose-induced enhancers showed enriched high-confidence (n = 439) or imputed (n = 640) assignments to glucose-induced genes, compared with distance-matched genes from the same TAD. Bottom: glucose-induced enhancers showed no enrichment for assignments to genes that were inhibited by high glucose concentrations (n = 196 interacting and n = 218 imputed pairs). OR = odds ratio. *P* values were calculated with Chi-square tests. e, Genes assigned to glucose-induced enhancers through high-confidence interactions (n = 275) or imputations (n = 321 pairs) were enriched for glucose-induced

promoter H3K27ac, compared with control genes from the same TAD. Box plots represent IQRs, notches are 95% confidence intervals of median, *P* values are from Wilcoxon's two-sided signed ranked tests. See also Supplementary Data Set 2.

We further tested enhancer-promoter assignments in a dynamic perturbation model. We exposed human islets from 7 donors to moderately low (4 mM) or high (11 mM) glucose 72 hours, which correspond to quasi-physiological glucose concentrations. This led to glucose-dependent H3K27ac changes in 3,850 enhancers at adjusted $P < 0.05$, most of which showed increased activity at high glucose (Supplementary Fig. 3o). This result, therefore, showed that changes in glucose concentrations elicit quantitative chromatin changes in a large number of human islet enhancers. We next reasoned that glucose-regulated enhancers should tend to cause glucose-regulated expression of their target genes. Indeed, we observed that glucose-induced enhancers were preferentially assigned to genes showing glucose-induced mRNA, compared with distance-matched active control genes from the same TAD (odds ratio 2.7 and 2.6, Fisher's $P = 4.9 \times 10^{-16}$ and 6.4×10^{-12}, for high-confidence or imputed assignments, respectively) (Fig. 2d). Likewise, genes assigned to glucose-induced enhancers showed greater glucose-induction of promoter H3K27ac than distance-matched promoters in the same TAD (Fig. 2e). Collectively, these studies validated pcHi-C maps for the identification of functional target genes of transcriptional enhancers in human pancreatic islets.

Genome editing of T2D-relevant enhancers.

A fundamental challenge to translate GWAS data into biological knowledge lies in identifying the target genes of noncoding elements that carry disease-associated regulatory variants. To link noncoding variants to their target genes, we compiled T2D- and fasting glycemia (FG)-associated variants from 109 loci, most of which have been fine-mapped to a credible variant set (Supplementary Fig. 4a, Supplementary Data Set 3). For fine-mapped loci, variants with a high posterior probability (PP > 0.1) of being causal were most enriched in active islet enhancers (Z = 20.9 relative to control regions in the same locus) and promoters (Z = 7.2)(Z < 2 for other accessible chromatin regions) (Supplementary Fig. 4b). In 61 loci we identified T2D- and/or FG-associated variants overlapping islet enhancers, and assigned one or more candidate target genes for 53 (87%) of these (Fig. 3a,

Supplementary Table 3). Some of these target genes were expected based on their linear proximity to the variants (e.g. *ADCY5*, *TCF7L2*, *ZFAND3*, *PROX1*, *FOXA2*), but for 75% of loci we identified more distant candidate genes. Examples of unexpected distal target genes, sometimes in addition to previously nominated proximal genes, include *SOX4* (in the *CDKAL1* locus), *OPTN* (*CDC123/CAMK1D*), *TRPM5* (*MIR4686*), *PDE8B* (*ZBED3*), *SLC36A4* (*MTNR1B*), *POLR3A and RPS24* (*ZMIZ1*), *MDGA1* (*ZFAND3*) and *PHF21A* (*CRY2*) (Fig. 3a, Supplementary Table 3, see *isletregulome.org* or *www.chicp.org*). Selected unexpected targets, including *ABCB9* and *STARD10*, were additionally supported by concordant eQTLs (Supplementary Fig. 4c-d).

We used genome editing to validate target genes of 10 enhancers bearing T2D- or FG-associated variants from 8 loci (Fig. 3b, Supplementary Table 4). We performed these experiments in EndoC-βH3 cells, a glucose-responsive human β cell line[20].

In the *CDC123 and CAMK1D* locus, only one SNP from a small set of fine-mapped T2D-associated variants is located in an islet enhancer (Fig. 3c, Supplementary Fig. 5a,b, Supplementary Table 3). This variant was previously proposed to be a regulatory variant based on plasmid reporter studies[21], allele-specific chromatin accessibility[22] and as an eQTL for *CAMK1D* [23,24] (Supplementary Table 2). The enhancer showed moderate-confidence interactions (CHiCAGO = 4.42) with *CAMK1D*, but, more surprisingly, showed high-confidence pcHi-C interactions with a more distant gene, *OPTN* (Fig. 3c, Supplementary Fig. 5a). Accordingly, deletion of this enhancer (but not an adjacent region), or silencing with KRAB-dCas9, led to selectively decreased expression of both *OPTN* and *CAMK1D*, whereas targeted activation of the enhancer stimulated their expression (Fig. 3d, Supplementary Fig. 5c,d). These results, therefore, confirm functional relationships predicted by pcHi-C maps. Although the role of *OPTN* and *CAMK1D* as mediators of this T2D-associated genetic signal remains to be defined, the findings highlight an example of a diabetes-relevant enhancer with multiple target genes.

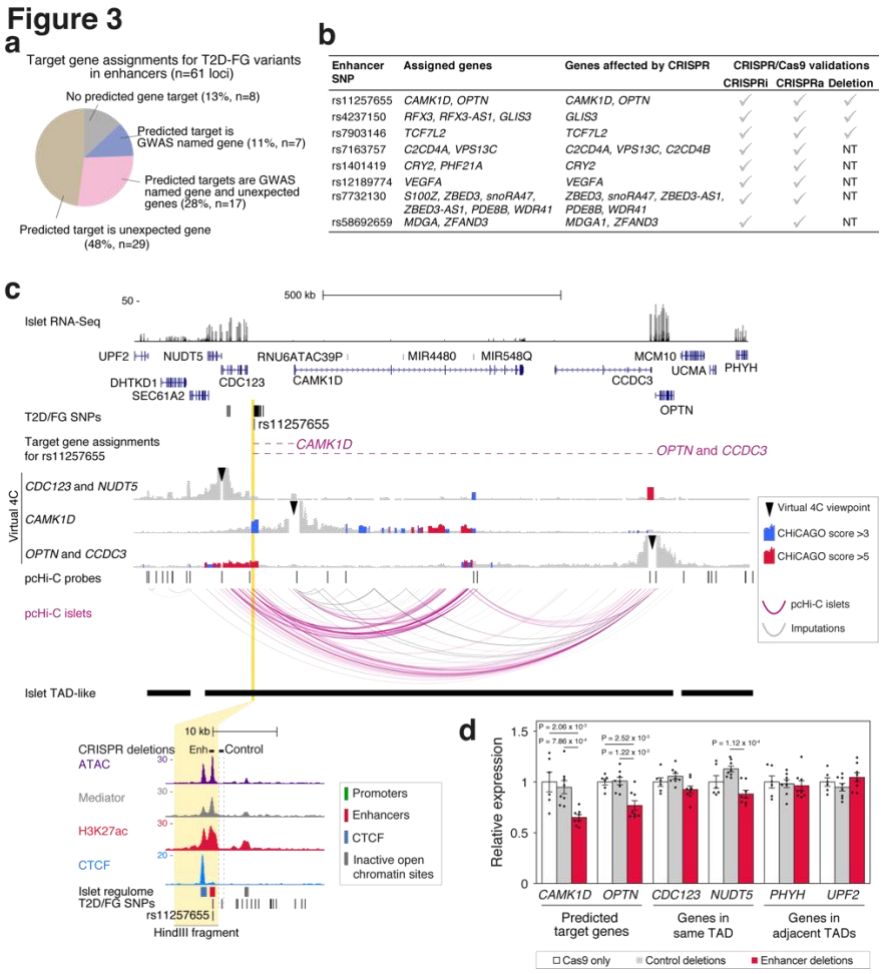


Figure 3. Identification of gene targets of T2D-relevant enhancers. **a,** We assigned gene targets through high-confidence interactions or imputations for 53 (87%) out of 61 T2D-FG associated loci with genetic variants in islet enhancers (Supplementary Table 3). **b,** Summary of T2D-associated enhancer perturbations presented in this study (see also Supplementary Table 4). NT, not tested. **c,** Islet pcHi-C analysis defines gene targets of enhancers bearing T2D-associated variants near *CDC123/CAMK1D*. The only T2D risk credible set variant that maps to an islet enhancer in the locus (rs11257655, zoomed inset) is assigned to *CAMK1D* and *OPTN* (dashed horizontal lines). Islet pcHi-C virtual 4C representations from pooled samples show interactions stemming from both *CAMK1D* and *OPTN* promoters towards rs11257655 with ChICAGO > 3, but not from *CDC123*. **d,** *CAMK1D* and *OPTN* mRNA are regulated by the rs11257655-containing enhancer. We deleted the rs11257655-containing enhancer and a nearby control region with a T2D-associated variant (rs33932777) that lacked active chromatin marks in human islets. Cas9 only: n = 6 (2 independent experiments with triplicates). Deletions: n = 8 (2 gRNA pairs in 2 independent experiments with biological duplicates). Bars are means ± s.e.m.,

normalized by *TBP* and expressed relative to mean levels of the Cas9 only controls. Statistical significance: two-tailed Student's *t* test.

We also examined rs7903146, a plausible causal SNP in the *TCF7L2* locus. This is the strongest known genetic signal for T2D, and it is known to influence islet-cell traits in non-diabetic individuals[2,25,26]. SNP rs7903146 lays in a class I enhancer with unusually high Mediator occupancy (Supplementary Fig. 6a). The SNP alters allele-specific accessibility and episomal enhancer activity[6], and has been associated with differences in *TCF7L2* mRNA[27]. However, deletion of this enhancer in human colon cancer cells affects *ACSL5* rather than *TCF7L2* [28], thereby questioning the true target genes of this enhancer in islet cells. We found that the rs7903146-bearing enhancer has imputed and moderate-confidence pcHi-C interactions with *TCF7L2*, but no evidence of proximity with any other gene in human islets (Supplementary Fig. 6a). Consistently, targeted deletion, functional inhibition, or stimulation of the enhancer caused selective changes in *TCF7L2* mRNA (Supplementary Fig. 6b,c). Therefore, the enhancer that harbors rs7903146 regulates *TCF7L2* in human β cells. Regardless of the possible metabolic role of this locus in other cell types[29], this finding indicates that *TCF7L2* is a likely mediator of the genetic association between rs7903146 and islet-related traits. For all 8 tested loci, at least one of the genes assigned by pcHi-C to an enhancer showed gene expression changes, and four showed changes in expression of more than one gene (Fig. 3b, Supplementary Table 4, Supplementary Data Set 4). This included functionally validated imputed target genes, such as *VEGFA* as well as *MDGA1* and *ZFAND3* (Supplementary Fig. 7). These functional studies, therefore, underscore the complexity of enhancer-promoter interactions, with long-range interactions that cannot be predicted from linear genome maps, interactions that are not functionally essential, and frequent target gene multiplicity. Importantly, the results validate the use of human pcHi-C maps to connect regulatory elements that harbor T2D-associated variants with the genes that can mediate disease susceptibility mechanisms.

Islet-specific transcription is linked to enhancer hubs.
Earlier studies demonstrated that risk variants for common diseases such as T2D are enriched in clusters of enhancers that regulate key

cell identity genes[4-7]. However, spatial clustering of enhancers is not necessarily apparent from linear genome maps. To identify 3D enhancer clusters, we again considered promoter-associated 3D spaces, or PATs, and empirically defined *enhancer-rich* PATs as those containing three or more class I enhancers (enhancers with high H3K27ac and Mediator occupancy, Fig. 1c). This definition of enhancer-rich PATs was supported by a multivariate analysis of genomic and epigenomic PAT features that were most predictive of islet-specific gene expression (Supplementary Fig. 8a and Methods). In total, we identified 2,623 enhancer-rich PATs (Supplementary Fig. 8b). As noted above, many active enhancers (~40%) had interactions with ≥1 promoter (Supplementary Fig. 3f). Thus, separate enhancer-rich PATs were often connected. We therefore merged enhancer-rich PATs with other PATs connected through enhancer-mediated high-confidence interactions, yielding 1,318 islet *enhancer hubs* (Fig. 4a, Supplementary Fig. 8c). Compared to alternate enhancer hub definitions, this definition maximized the enrichment of islet cell functional annotations and the number of mapped hubs (Supplementary Fig. 9). The 1,318 islet enhancer hubs are, in essence, 3D chromatin domains that contain a median of 18 enhancers, two active promoters, and two shared enhancer interactions (Supplementary Fig. 8d). They are often tissue-selective interaction domains, because hub promoters had 2.8-fold higher fraction of islet-selective interactions than non-hub promoters (Wilcoxon's $P = 2.8 \times 10^{-36}$) (Supplementary Fig. 8e, examples in Figs. 1b, 5a, Supplementary Figs. 1h,i and 10a). Furthermore, the genes that form part of enhancer hubs were enriched in islet-selective transcripts, and in functional annotations that are central to islet cell identity, differentiation, and diabetes (Fig. 4b,c, Supplementary Table 5, Supplementary Data Set 5).

Hubs exhibit domain-level chromatin changes.
Consistent with the high internal connectivity of hubs, gene pairs from the same hub showed increased RNA expression correlation values across tissues and islet samples, as compared to control active gene pairs in the same TAD as the hubs ($P = 6.3 \times 10^{-8}$)(Fig. 4d). Moreover, hub enhancers showed higher H3K27ac correlations with their target promoters than when were paired with non-hub promoters from the same TAD ($P = 2.2 \times 10^{-16}$) (Fig. 4e). These findings are consistent with enhancer interaction hubs as functional regulatory domains.

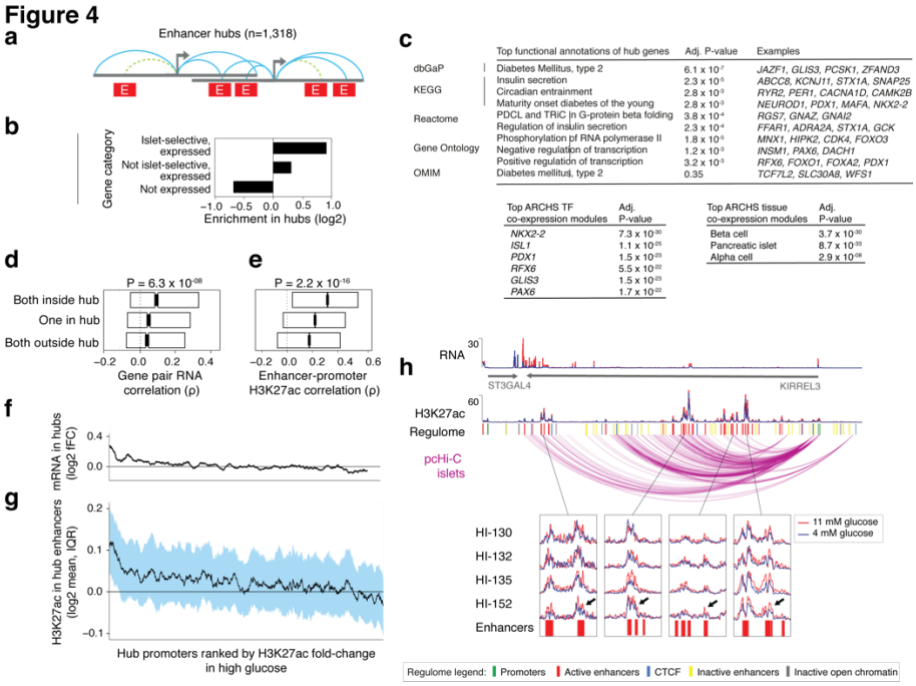


Figure 4. Tissue-specific enhancer hubs regulate key islet genes.
a, Hubs are composed of one or more enhancer-rich PATs (≥ 3 class I enhancers) connected through at least one common interacting enhancer. Turquoise and dashed green lines depict high-confidence and imputed assignments, respectively. Descriptive features of hubs are summarized in Supplementary Figure 8c. **b,** Islet hubs are enriched in genes showing islet-selective expression. Ratios were calculated relative to all annotated genes. **c,** Islet hub genes are enriched in annotations important for islet differentiation, function and diabetes. Benjamini-Hochberg adjusted *P* values from EnrichR are shown (see complete lists in Supplementary Table 5). **d,** Gene pairs from the same hub show higher RNA correlations across human islet samples and 15 control tissues than gene pairs from the same TAD in which only one gene or neither gene is in a hub. *P* values were derived with Kruskall-Wallis analysis of variance. **e,** Enhancer-promoter pairs from the same hub show high H3K27ac correlations across 14 human islet samples and 51 Epigenome RoadMap tissues, compared with pairs from the same TAD in which only one element or neither are in a hub. *P* values were derived with a Kruskall-Wallis test. **f,-g,** Culture of 7 human islet donor samples at 4 vs. 11 mM glucose shows concerted changes in H3K27ac in hub enhancers connected with glucose-dependent genes. Hub promoters were ranked by their median fold-change in H3K27ac at high glucose, so that glucose-induced promoters are on the left of the X axis. (f) Median mRNA for genes associated with each hub. (g) Median glucose-dependent fold-change of H3K27ac in enhancers from hubs connecting with each promoter, IQR values in blue shade. In both graphs values are shown as running averages (window = 50). h, Coordinated glucose-induced H3K27ac in enhancers of a hub connected to *KIRREL3*. Top tracks show RNA and H2K27ac in one representative sample.

Bottom insets highlight H2K27ac at 11 mM glucose (red) vs. 4 mM (blue) in regions showing coordinated glucose-induced changes in most hub enhancers, highlighted with black arrows (n = 4 human islet samples). See also Supplementary Table 6, Supplementary Data Set 5.

To further explore the behavior of hubs as functional domains, we again examined islets exposed to moderately low vs. high glucose concentrations. Glucose-induced enhancers and mRNAs were highly enriched in hubs, compared with non-hub counterparts (Fisher's $P = 1.1 \times 10^{-7}$ and 2.2×10^{-16}, respectively). Of 297 promoters that showed glucose-induced H3K27ac, 94 were contained in hubs, and 65% of these showed glucose-induced mRNA (Supplementary Tables 6,7). We predicted that if hubs are functional regulatory domains, hub enhancers connected to glucose-induced genes should tend to show coordinated glucose-dependent changes. Our analysis showed that hub enhancers assigned to glucose-induced promoters showed a widespread parallel increase in H3K27ac (Fig. 4f-h, Supplementary Table 8). Thus, varying glucose concentrations elicit chromatin changes in human islets at the level of broad regulatory domains. Taken together, our findings indicate that enhancer hubs have properties of functional units.

Enhancer hubs contain super-enhancers and enhancer clusters.
We compared islet enhancer hubs with previously defined islet enhancer domains, such as linear enhancer clusters and super-enhancers (Supplementary Fig. 8f). This showed that hubs have at least some spatial overlap with 70% of enhancer clusters [7], and with 87% of super-enhancers defined with a standard algorithm [4] (Supplementary Fig. 8g-i). Hubs, however, differ in that they can be connected with their target genes. Furthermore, enhancer hubs capture spatial clusters of Mediator-bound (class I) enhancers that do not cluster in the linear genome and therefore do not fulfill definitions of super-enhancers and enhancer clusters (Supplementary Fig. 8j-l) [4,7]. In fact, many hubs contained several inter-connected enhancer clusters or super-enhancers (Supplementary Fig. 8m-o). This is illustrated by the *ISL1* locus, which has several enhancer clusters and super-enhancers distributed across an entire TAD, whereas pcHi-C points to a single hub that connects dozens of enhancers with *ISL1* and lncRNA *HI-LNC57* (Fig. 5a). Thus, enhancer hubs are 3D domains that often include

one or more enhancer clusters or super-enhancers and their target gene(s).

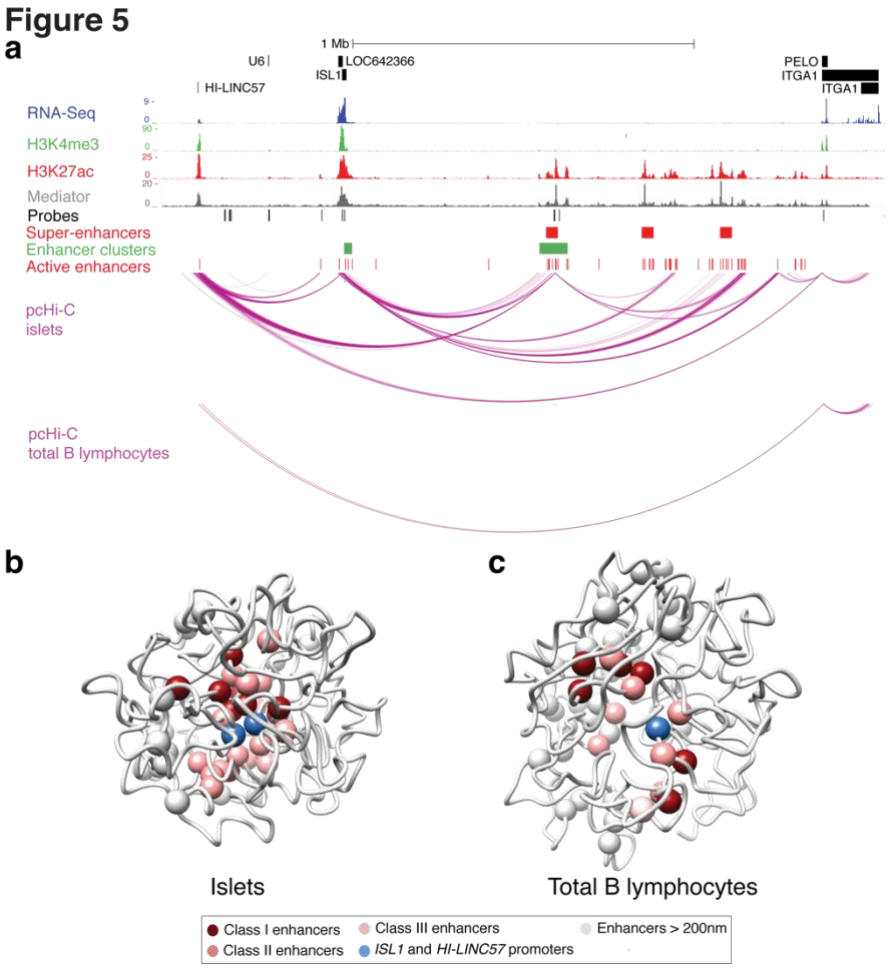


Figure 5. Tissue-specific topology of the *ISL1* enhancer hub. a, Epigenomic annotations and high-confidence pcHi-C interactions from pooled islet samples and total B lymphocytes are shown to illustrate active enhancers, super-enhancers and enhancer clusters distributed across a TAD, while sharing islet-selective 3D interactions with *ISL1* and *HI-LNC57*. **b-c**, 3D chromatin conformation models of the *ISL1* enhancer hub generated from pcHi-C libraries from human islets (**b**) and total B lymphocytes (**c**). Images represent the top scoring model from the ensemble of structures that best satisfied spatial restraints. Class I, II and III enhancers are colored in dark to light red and promoters in blue if they are within 200 nm of the *ISL1* promoter, or as white spheres if they are further than 200 nm. Note the proximity of lncRNA *HI-LNC57* and *ISL1* promoters in islets. The models show that active islet regulatory elements interact in a restricted 3D space in islet nuclei. See also Supplementary Figure 10b,c and Supplementary Videos 1 and 2**.**

Tissue-specific architecture of the *ISL1* enhancer hub.
To gain insight into the 3D conformation of enhancer hubs, we built 3D models of hubs using islet pcHi-C interaction data (Fig. 5a). We focused on the *ISL1* locus because it contains a single hub within a TAD-like domain, with few other annotated genes. We used islet pcHiC data to build interaction matrices at 5-kb resolution, and transformed the frequency of interactions between genomic segments into spatial restraints [30,31]. We then used molecular dynamic optimization to generate an ensemble of 500 models that best satisfied the imposed restraints. This showed co-localization of islet enhancers and target genes in a constrained space of the TAD, whereas models built from B lymphocyte pcHi-C libraries showed decreased aggregation of these regions (Fig. 5b,c, Supplementary Fig. 10b,c, Supplementary Videos 1-2). Quantitative analysis of *ISL1* and six other T2D-relevant hubs showed analogous tissue-specific aggregation of hub enhancers and promoters (Supplementary Figs. 10d-I, 13f-h). These models, which capture the average topology in a population of cells, serve to highlight that whereas TADs are defined as single intervals in linear genome maps, hubs are formed by multiple interspersed regions that occupy a shared 3D subspace within a TAD.

Epigenome editing of T2D-associated islet hubs.
We used enhancer perturbations to test the functional connectivity of selected enhancer hubs. In the *ZBED3* locus, we targeted a class I enhancer that contains a variant with highest posterior probability for causality in T2D fine-mapping studies (PP = 0.461) (Fig. 6a, Supplementary Fig. 11a, Supplementary Table 4). Targeted epigenomic activation or inhibition of this single enhancer led to significant changes in the expression of five of the six genes connected with this hub, but not of non-hub genes from the same TAD (Fig. 6b). In three other hubs we perturbed single enhancers containing candidate T2D susceptibility causal variants, which led to expression changes in *CRY2* and *PHF21A* (Supplementary Fig. 11b,c), *VPS13C, C2CD4A* and *C2CD4B* (Supplementary Fig. 12) and *GLIS3* (Supplementary Fig. 13). These findings highlight a remarkable functional connectivity of enhancer hubs.

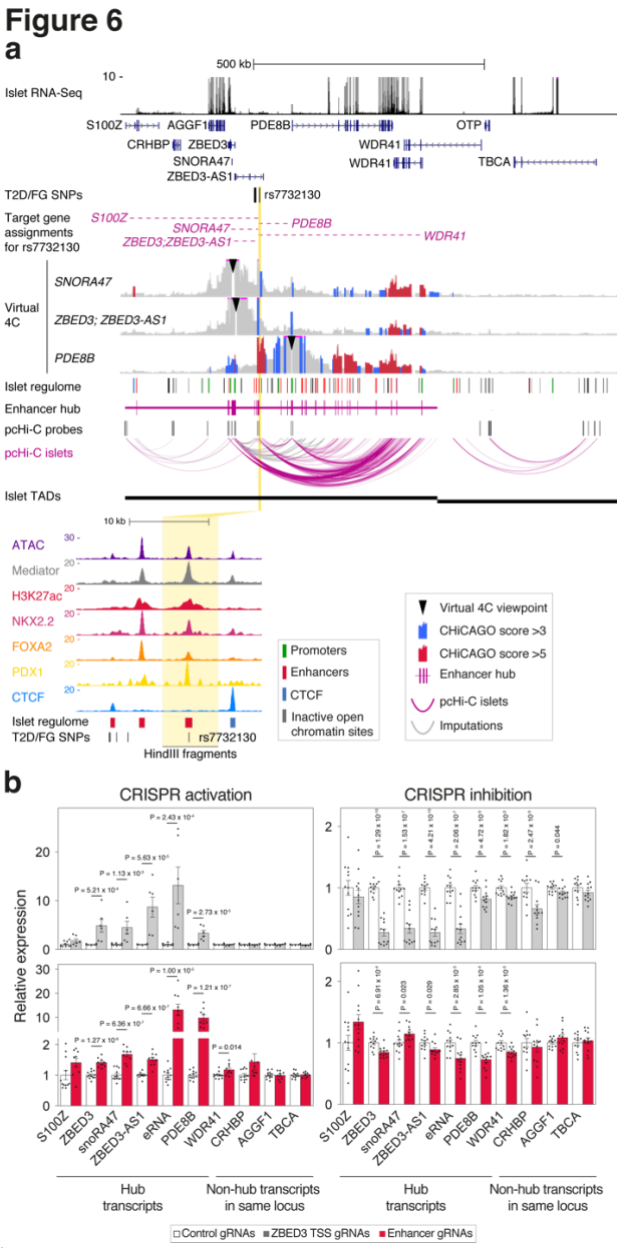


Figure 6. The *ZBED3* enhancer hub links an enhancer bearing a T2D SNP with multiple target genes. a, pcHi-C and virtual 4C representations from pooled islet samples for three viewpoints (see also Supplementary Fig. 10). The variant with highest posterior probability in this locus (rs7732130) maps to a class I islet enhancer (yellow line, and zoomed inset) that shows interactions with *PDE8B* (CHiCAGO > 5), and *ZBED3, ZBED3-AS1*, *snoRA47* and *S100Z* (CHiCAGO > 3, see also Supplementary Fig. 11). WDR41 is assigned to rs7732130 by imputation. Dashed horizontal lines show all targets assigned through imputation or high-confidence interactions. **b**, Analysis of hub and non-hub transcripts after CRISPR activation or

inhibition of the transcriptional start site of *ZBED3* or the rs7732130-enhancer in EndoC-βH3 cells. Data are presented as means ± s.e.m. of all gRNAs combined per target region (enhancer CRISPRa: 3 gRNAs, CRISPRi: 4 gRNAs, all n = 3 independent experiments). Statistical significance: two-tailed Student's *t* test.

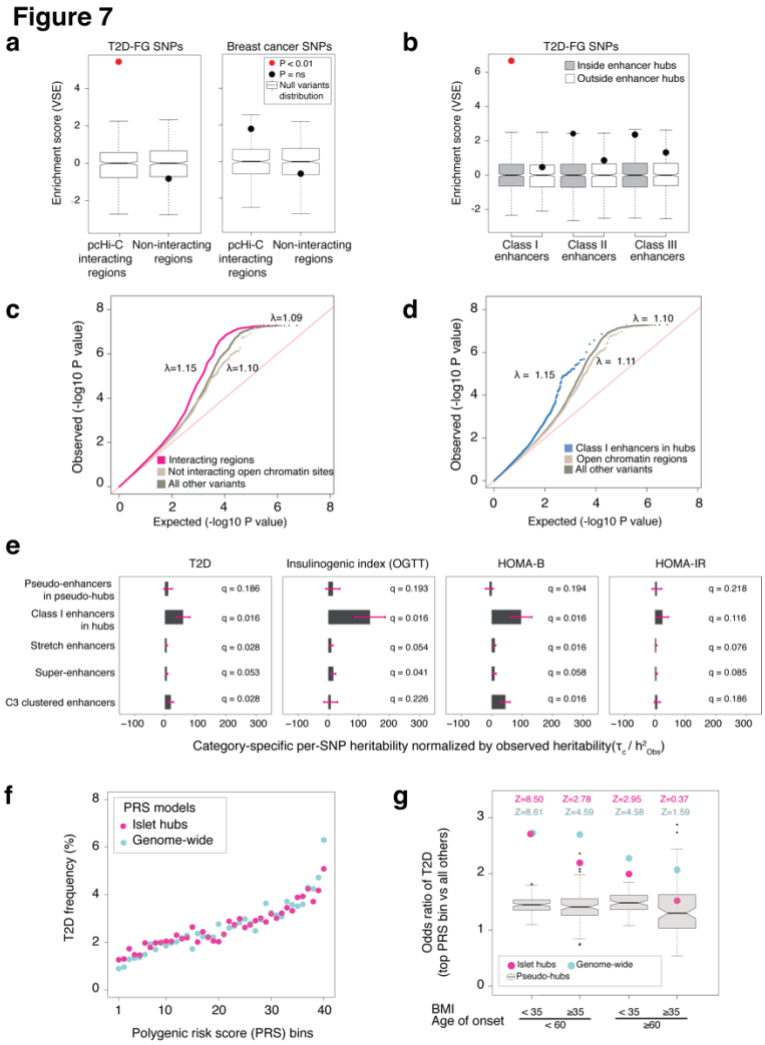


Figure 7. Islet hub variants impact insulin secretion and provide tissue-specific risk scores. **a**, Variant Set Enrichment (VSE) for T2D and FG (n = 2,771 variants; Supplementary Table 9) and breast cancer (n = 3,048 variants) in high-confidence islet interacting fragments. Box plots show 500 permutations of matched random haplotype blocks. Red dots indicate significant enrichments (Bonferroni–adjusted $P < 0.01$). **b**, T2D and FG GWAS significant variants are selectively enriched in hub class I islet enhancers. Boxplots show median and IQR. **c**, Genomic inflation of T2D association *P* values for non-GWAS significant variants (P > 5 ×

10^{-8}) from a T2D GWAS meta-analysis (12,931 cases, 57,196 controls) in islet high-confidence interacting regions (magenta), non-interacting islet open chromatin (beige), and all other variants (brown). **d**, Genomic inflation of T2D association *P* values for non-GWAS significant variants in hub class I islet enhancers (blue), non-hub islet open chromatin (beige) and all other variants (brown). **e**, Heritability estimates based on GWAS summary statistics for T2D (12,931 cases, 57,196 controls), insulinogenic index (OGTT, 7,807 individuals), homeostasis model assessment of β-cell function (HOMA-B) and insulin resistance (HOMA-IR) (~80,000 individuals), for indicated islet enhancer domains. Bars show category-specific per-SNP heritability coefficients (τ_c) divided by LD score heritability (h^2) of each trait. τ_c coefficients were obtained independently for each trait, controlling for 53 functional annotation categories. Values were multiplied by 10^7 and shown with s.e.m. **f,** T2D frequency across 40 bins, each representing 2.5% of individuals in the UK Biobank test dataset (226,777 controls, 6,127 T2D cases) with increasing PRS, calculated with hub (pink dots) or genome-wide variants (light green). **g,** Odds ratios (OR) for T2D calculated for 2.5% individuals with highest PRS vs. all other individuals, using islet hub (pink) or genome-wide models (green), stratified by BMI and T2D age of onset. Boxplots show ORs for PRS from 100 permutations of pseudo-hubs (IQRs). Z-scores are standard deviations of pseudo-hub averages. See also Supplementary Figure 15 and Supplementary Table 17.

Islet hub variants impact insulin secretion.

Previous evidence that T2D susceptibility variants are enriched in islet enhancer clusters [5-7,24,32] prompted us to examine the enrichment of diabetes-associated variants in our newly defined annotations. T2D/FG-associated SNPs were enriched in islet pcHi-C interaction regions (Fig. 7a), and in islet enhancer hub class I enhancers, rather than in other active enhancers (Fig. 7b, Supplementary Fig. 9, 14a-f, Supplementary Table 9). This indicates that hub class I enhancer variants are important for T2D susceptibility.

A major portion of the heritability of common diseases is driven by many variants that individually have not achieved genome-wide significance, yet exert a large aggregate effect [33-35]. Consistent with this notion, common variants that have so far not shown genome-wide significance for T2D association, but are located in pcHi-C interacting regions or hub class I enhancers, showed more significant association *P* values than expected distributions (Fig. 7c,d). This observation prompted us to quantify the overall contribution of common variants in islet hubs to the heritability of T2D. We used stratified LD score regression [36], and found that hub class I enhancers showed the most significantly increased per-SNP T2D heritability coefficient ($q = 1.64 \times 10^{-2}$) compared with various

islet and non-islet genomic annotations (Fig. 7e, Supplementary Fig. 15a, Supplementary Table 10).

Although islet dysfunction is central to the pathophysiology of T2D, other tissues (liver, adipose, muscle, brain, among others) are also critically important[37]. Genetic variation in islet hub enhancers should, therefore, predominantly impact on the heritability of pancreatic islet function. Indeed, islet hub variants showed higher heritability enrichment estimates for islet-cell traits than for T2D (Fig. 7e, Supplementary Fig. 15a-f, Supplementary Table 10). Consequently, common variation in hub class I enhancers (0.26% of genomic SNPs) explained 9.9% of observed genetic heritability for T2D, 21.9% for acute insulin secretory response in intravenous glucose tolerance tests[26], 17.2% for HOMA-B models of β-cell function, and 31.2% for an insulinogenic index based on oral glucose tolerance tests[38] (Supplementary Table 10). In sharp contrast, islet hub variants showed no enrichment for HOMA-IR, an estimate of insulin resistance (Supplementary Fig. 15e). Of note, significant heritability enrichments were generally also observed for enhancer clusters, stretch enhancers, or super-enhancer annotations, yet estimates were consistently larger for hub enhancers (Fig. 7e, Supplementary Fig. 15a-d). These results indicate that enhancer hubs define genomic spaces that play a prominent role in the heritability of T2D and insulin secretion.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0457-0.

References

1. Chatterjee, S., Khunti, K. & Davies, M.J. Type 2 diabetes. *Lancet* **389**, 2239-2251 (2017).
2. Flannick, J. & Florez, J.C. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat Rev Genet* **17**, 535-49 (2016).
3. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41-47 (2016).

4. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-19 (2013).
5. Parker, S.C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A* **110**, 17921-6 (2013).
6. Gaulton, K.J. *et al.* A map of open chromatin in human pancreatic islets. *Nat Genet* **42**, 255-9 (2010).
7. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* **46**, 136-143 (2014).
8. Cohen, A.J. *et al.* Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. *Nat Commun* **8**, 14400 (2017).
9. Farh, K.K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-43 (2015).
10. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).
11. Vahedi, G. *et al.* Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature* **520**, 558-62 (2015).
12. Montavon, T. *et al.* A regulatory archipelago controls Hox genes transcription in digits. *Cell* **147**, 1132-45 (2011).
13. Patrinos, G.P. *et al.* Multiple interactions between regulatory regions are required to stabilize an active chromatin hub. *Genes Dev* **18**, 1495-509 (2004).
14. Javierre, B.M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384 e19 (2016).
15. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol* **17**, 127 (2016).
16. Schofield, E.C. *et al.* CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. *Bioinformatics* **32**, 2511-3 (2016).
17. Mularoni, L., Ramos-Rodriguez, M. & Pasquali, L. The Pancreatic Islet Regulome Browser. *Front Genet* **8**, 13 (2017).

18. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).
19. Nora, E.P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-5 (2012).
20. Benazra, M. *et al.* A human beta cell line with drug inducible excision of immortalizing transgenes. *Mol Metab* **4**, 916-25 (2015).
21. Fogarty, M.P., Cannon, M.E., Vadlamudi, S., Gaulton, K.J. & Mohlke, K.L. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. *PLoS Genet* **10**, e1004633 (2014).
22. Thurner, M. *et al.* Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes susceptibility loci. *Elife* **7**(2018).
23. van de Bunt, M. *et al.* Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. *PLoS Genet* **11**, e1005694 (2015).
24. Varshney, A. *et al.* Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc Natl Acad Sci U S A* **114**, 2301-2306 (2017).
25. Scott, R.A. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* **66**, 2888-2902 (2017).
26. Wood, A.R. *et al.* A Genome-Wide Association Study of IVGTT-Based Measures of First-Phase Insulin Secretion Refines the Underlying Physiology of Type 2 Diabetes Variants. *Diabetes* **66**, 2296-2309 (2017).
27. Lyssenko, V. *et al.* Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *J Clin Invest* **117**, 2155-63 (2007).
28. Xia, Q. *et al.* The type 2 diabetes presumed causal variant within TCF7L2 resides in an element that controls the expression of ACSL5. *Diabetologia* **59**, 2360-2368 (2016).
29. Nobrega, M.A. TCF7L2 and glucose metabolism: time to look beyond the pancreas. *Diabetes* **62**, 706-8 (2013).

30. Bau, D. *et al.* The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* **18**, 107-14 (2011).
31. Serra, F. *et al.* Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput Biol* **13**, e1005665 (2017).
32. Gaulton, K.J. *et al.* Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat Genet* **47**, 1415-25 (2015).
33. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186 (2017).
34. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-86 (2014).
35. Khera, A.V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219-1224 (2018).
36. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).
37. DeFronzo, R.A. *et al.* Type 2 diabetes mellitus. *Nat Rev Dis Primers* **1**, 15019 (2015).
38. Gjesing, A.P. *et al.* Genetic and phenotypic correlations between surrogate measures of insulin release obtained from OGTT data. *Diabetologia* **58**, 1006-12 (2015).
39. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet* **50**, 1505-1513 (2018).
40. Khera, A.V. *et al.* Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell* **177**, 587-596 e9 (2019).
41. Richardson, T.G., Harrison, S., Hemani, G. & Davey Smith, G. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife* **8**(2019).
42. Bonas-Guarch, S. *et al.* Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. *Nat Commun* **9**, 321 (2018).

43. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
44. Bycroft, C.F., D.; Petkova, G.; Band, L.T.; Elliott, K.; Sharp, A.; Motyer, D.; Vukcevic, O.; Delaneau, J.; O'Connell, A.; Cortes, S.; Welsh, G.; McVean, S.; Leslie, P.; Donnelly, J.; Marchini. Genome-Wide Genetic Data on ~ 500,000 UK Biobank Participants. *BioRxiv* (2017).
45. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
46. Schmitt, A.D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep* **17**, 2042-2059 (2016).
47. Harmston, N. *et al.* Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat Commun* **8**, 441 (2017).
48. Akalin, A. *et al.* Transcriptional features of genomic regulatory blocks. *Genome Biol* **10**, R38 (2009).
49. Ahlqvist, E. *et al.* Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol* (2018).
50. Kahn, S.E., Cooper, M.E. & Del Prato, S. Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. *Lancet* **383**, 1068-83 (2014).

Acknowledgements
This research was supported by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre. Work was funded by grants from the Wellcome Trust (WT101033 to J.F. and WT205915 to I.P.), Horizon 2020 (Research and Innovation Programme 667191 to J.F., 633595 to I.P. and 676556 to M.A.M.-R.; Marie Sklodowska-Curie 658145 to I.M.-E., and 43062 ZENCODE to G.A.), European Research Council (789055 to J.F., 609989 to M.A.M.-R.). Marató TV3 (201611, to J.F., M.A.M.-R.), Ministerio de Ciencia Innovación y Universidades (BFU2014-54284-R, RTI2018-095666 to J.F., BFU2013-47736-P to M.A.M.-R., IJCI-2015-23352 to I.F.), AGAUR (to M.A.M.-R.). UK Medical Research Council (MR/L007150/1 to P.F.), World Cancer Research Fund (WCRF UK to I.P.) and World Cancer Research Fund

International (2017/1641), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL, NWO 184.021.007 to I.O.F.). Work in IDIBAPS, CRG and CNAG was supported by the CERCA Programme, Generalitat de Catalunya and Centros de Excelencia Severo Ochoa (SEV-2012-0208). Human islets were provided through the European islet distribution program for basic research supported by JDRF (3-RSC-2016-160-I-X). We thank Natalia Ruiz-Gomez for technical assistance, Rodrigo Liberal Fernandes, Thomas Thorne (University of Reading), and Alvaro Perdones-Montero (ICL) for helpful discussions regarding Machine Learning approaches, Boris Lenhard and Matthias Merkenschlager (LMS, ICL), Ferenc Müller (UoB) and José Luis Gómez-Skarmeta (CABD) for critical comments on the draft; the CRG Genomics Unit, and the Imperial College High Performance Computing Service.

Author contributions

I.M.-E., I.C., and B.M.J. performed and analyzed experiments. I.M.-E. and J.G.-H. processed human islet samples. I.M.-E., S.B.-G., I.C., J.P.-C., D.M.Y.R., G.A., C.C.M. and I.M. performed computational analysis. J.M.-E. and I.F. modeled and analyzed 3D data. L.Pi., T.B., E.J.P.d.K., J.K.-C., F.P. and P.R. provided material and reagents. E.V.R.A., A.L., A.P.G., D.R.W., O.P., N.G., J.M.M., D.T., I.O.F., I.P., and L.G. provided genetics data. M.R. and L.Pa. created software resources. I.C. and A.B. developed genome-editing methods. M.M.-R., P.F. and J.F. supervised analysis. I.M.-E., I.C., S.B-G., J.P-C., D.M.Y.R. and J.F. conceived the project. I.M-E., S.B-G., I.C., and J.F. wrote and edited the manuscript, which all authors have approved.

Competing Interests Statement

P.R. is a shareholder and consultant for Endocells/Unicercell Biosolutions.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/ s41588-019-0457-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Human islets. Human pancreatic islets from organ donors without a history of glucose intolerance were purified using established isolation procedures[1-4], shipped in culture medium and re-cultured at 37°C in a humidified chamber with 5% CO_2 in glucose-free RPMI 1640 supplemented with 10% fetal calf serum, 100 U/ml penicillin, 100 U/ml streptomycin and 11 mM glucose for three days before analysis. RNA was extracted from flash-frozen islet pellets using TRIrizol Reagent (Thermo Scientific). For glucose regulation studies, islets were cultured in identical time and medium, except that glucose-free RPMI 1640 medium was supplemented with glucose to achieve final concentrations of either 4 or 11 mM glucose. Donor and sample characteristics are provided in Supplementary Table 11.
Compliance with ethic regulations for human research studies is described in Supplementary Note 1.

pcHi-C. 30-60 million human islet cells/donor from four islet donors were cultured as described above for three days prior to fixation in 2% paraformaldehyde (Agar Scientific) at room temperature for 10 minutes with mixing, quenched in 125 mM glycine for 5 minutes at room temperature and 15 minutes in ice, and washed twice in PBS. Dry pellets were flash frozen and stored at -80°C.

Hi-C libraries were prepared with in-nucleus ligation and processed to capture 22,076 HindIII fragments containing 31,253 annotated promoters for 18,202 protein-coding and 10,929 non-protein coding genes (Ensembl v.75), using SureSelect target enrichment (Agilent Technologies), as described previously[5,6]. After library enrichment, a post-capture PCR amplification step was carried out with 4 PCR amplification cycles.

Twelve sequencing replicates from 4 human islet donor libraries were processed using a reported pipeline which maps di-tags against the human genome (GRCh37), filters out experimental artefacts such as re-ligations, and removes PCR duplicates[7]. Reads

from replicates from each donor were then pooled. Alignment statistics are shown in Supplementary Tables 12,13.

Interaction confidence scores were computed with CHiCAGO[6,8]. High confidence interactions were defined as CHiCAGO scores >5, as described[6]. pcHi-C datasets from unrelated tissues[6] were processed identically. CHiCAGO analysis is generally performed with pooled libraries as this increases sensitivity and mitigates subsampling in individual libraries[6,8]. We assessed reproducibility across individual samples, and observed that high-confidence interaction calls showed (a) high CHiCAGO scores in individual samples, with limited overlap with distance-matched regions (Supplementary Fig. 1d), (b) pairwise Pearson ρ values of individual sample CHiCAGO scores ranging 0.62-0.74 (Supplementary Fig. 1e), (c) consistent above-background scores in individual samples (Supplementary Figs. 1f,g and 5a).

ChIP-seq and ATAC-seq. ChIP and ATAC were performed as previously described[9,10], with modifications (Supplementary Note 2). Adaptor trimming of ChIP-seq reads was performed with cutadapt 1.9.1 (options: -m 20)[11]. For ATAC-seq, low quality bases and adaptor trimming were processed using Trimgalore 0.4.1 (options --quality 15 –nextera). Trimmed reads were aligned to hg19 using bowtie2 2.1.0 (options: --no-unal) allowing no mismatches[12], retaining uniquely mapped reads (MAPQ>=30) using samtools 1.2 [13], removing duplicate reads (picard 2.6.0)[14], blacklisted regions[15], and, for ATAC-seq, mitochondrial reads. Data quality was assessed with SPP.R script from phantompeaktools[16]. ChIP-seq and ATAC-seq information is shown in Supplementary Table 1.

For histone modifications, broad enriched regions were called with MACS2[17] using --g hs --extsize=300 --keep-dup all --nomodel --broad and narrow regions were called without using --broad flag. For TF and co-factors, narrow regions were called using –g hs –extsize=300 --keep-dup all. For ATAC-seq, we used --shift 100 --extsize=200 --keep-dup all --nomodel.

To obtain a robust set of ChiP-seq peaks, we called peaks in individual human islet samples with relaxed stringency ($P < 0.01$), and in pooled samples using a stringent threshold (FDR $q < 0.05$ for

Mediator and cohesin; and q < 0.01 for histone modification marks). We then identified peaks present in at least 3 individual samples, or at least 2 samples if only 3 replicates were processed, as well as in the pooled set. For accessible chromatin sites, we called peaks at *P* < 0.01 in 13 individual samples, and FDR q < 0.05 from pooled samples. We then defined consistent peaks present in at least 3 samples as well as in the pooled set. Consistent ATAC peaks that showed multiple sub-peaks in > 3 islet samples were manually split, leading to n = 241,481 ATAC peaks. A final set of accessible chromatin regions (n = 249,582) was defined by adding regions lacking ATAC-seq peaks that showed either Mediator or CTCF binding (n = 1,319, n = 9,596 respectively) or were bound by at least two islet transcription factors (n = 1,514)[9]. bigwig files were generated using *bamCoverage* from deepTools (-e=300 --normalizeTo1x 2451960000).

Classification of human islet accessible chromatin. We classified 249,582 consistent islet open chromatin regions using k-medians clustering of ChIP-seq signal distribution of H3K27ac, H3K4me1, H3K4me3, Mediator, cohesin and CTCF, using islet samples with greatest signal to noise for these marks. Briefly, $-\log_{10}$ (*P* value) signal was calculated for each mark using 100 bp bins across a 6-kb window centered on consistent open chromatin regions. K-median clustering (flexClust [18]) was used to classify open chromatin regions into 14 clusters, which were manually merged into 8 clusters based on the chromatin mark enrichment patterns. Each open chromatin class was ranked by CTCF binding to highlight a subset of CTCF-bound enhancers. Post-hoc analysis showed that human islet transcription start sites defined by CAGE were markedly enriched in regions classified as active promoters, and to a lesser extent in class I enhancers (Fig. 1c). See Supplementary Data Set 1 for genomic locations.

PATs and enhancer-promoter assignments. We defined 16,030 Promoter-Associated Territories (PATs) as the linear space covered by all interactions originating from a pcHi-C bait, within the same islet TAD-like compartment (Supplementary Note 3).

We used PAT features to assign enhancers to promoters, following a stepwise approach such that each step was performed on unassigned enhancers from previous steps. We assigned enhancers

to baits with at least one active islet promoter according to our regulome annotations (Supplementary Data Set 1) or ChromHMM analyses (Supplementary Note 6), and report target genes with average human islet RNA expression > 1.5 TPM (Supplementary Data Set 2), based on the following criteria:
1. Presence of high-confidence interactions (CHiCAGO score > 5) to one or more baits, including those that cross TAD boundaries (also referred to as assignment by interaction).
2. For enhancers with no high-confidence interactions, we defined PAT(s) in which they were contained. We did not assign enhancers to all overlapping PATs because only some active genes are regulated by enhancers, and instead only imputed orphan enhancers to PAT(s) anchored by an active promoter that already showed high-confidence interactions with other islet enhancers.
3. For remaining enhancers located < 10 kb away from a bait containing active promoter(s), we assumed that (a) this linear distance is more likely to provide functional enhancer-promoter communication than promoters located more distally that do not show high-confidence interactions, and (b) random collisions are too frequent to detect high confidence interactions above background noise, and thus imputed these enhancer-promoter assignments.
4. For remaining enhancers that were exclusively contained within a single PAT with an active promoter, we imputed the assignment to expressed genes in that PAT bait. We refer to assignment criteria 2-4 as imputations in the manuscript.
Enhancer-promoter assignments can be found in Supplementary Data Set 2, and were validated by analysis of (a) CHiCAGO scores in imputations (b) increased enhancer-promoter correlations, (c) islet-specificity of assigned genes, (d) concordance with eQTLs, and (e) coordinated changes after exposure to varying glucose concentrations (Supplementary Note 7).

Candidate target genes of T2D-FG associated variants. We integrated lists of T2D/FG-associated variants (Supplementary Note 8) with enhancer-promoter assignments to identify candidate target genes. We associated 555 enhancer variants from 51 loci to islet-expressed genes using high-confidence interactions and imputations. Supplementary Table 3 provides a more extensive list of 830 T2D/FG-associated variants overlapping an active enhancer or promoter, with information on connections to candidate target

genes through (a) high-confidence interactions (CHiCAGO score > 5), (b) moderate-confidence interactions (CHiCAGO 2.5-5), (c) imputations, (d) indirect connections through a common hub, and (e) location of actively expressed gene within 10 kb. This category also included actively transcribed genes from associated variant-containing promoters that overlap pcHi-C baits. Supplementary Table 3 additionally lists T2D-FG variants overlapping a promoter interacting region that do not overlap an annotated regulatory element.

Cell-based genome and epigenome editing. Experimental validation of T2D-relevant enhancer-promoter assignments in EndoC βH3 cells[19] is described in the Supplementary Note 10, and Nature Protocol Exchange (I.C. and Anthony Beucher).

Classification of PATs based on enhancer content.
We defined enhancer-rich PATs as those with three or more class I enhancers (Supplementary Fig. 8b). This was supported by logistic regression analysis (Supplementary Note 11) showing that the number of class I enhancers assigned to a PAT was independently predictive of islet-selective expression of PAT genes. This effect was optimized with PATs with ≥ 3 assigned class I enhancers (Supplementary Fig. 9).

Enhancer hubs. Enhancer-rich PATs were frequently interconnected through one or more shared enhancers (42.4% of all active enhancers had high-confidence interactions with > 1 bait). We thus merged enhancer-rich PATs with other PATs that were connected by one or more common enhancers through high-confidence interactions (CHiCAGO score > 5). For 99.5% of hubs all hub components were restricted to one chromosome. Alternative definitions of hubs were created to test how a) the number of enhancers in enhancer-rich PATs, b) the inclusion of enhancer-gene imputed assignments, and c) criteria to merge PATs, influence definitions of enhancer hubs (Supplementary Fig. 9).

To annotate hub genes, we considered annotated promoters of genes with median RNA expression > 1.5 TPM in human islets. In few cases (n = 426), pcHi-C bait fragments contained active enhancers that established high-confidence pcHi-C interactions with non-baited fragments containing active islet promoters, which were also

considered as constituents of islet hubs. A list of human islet enhancer hubs is presented in Supplementary Data Set 5. Functional enrichments of hub Ensembl genes were performed with Enrichr[20]. The analysis of correlated hub promoter and enhancer activity, and islet-selectivity of enhancer interactions is described in Supplementary Note 12.

3D modeling of hubs. 3D modeling and analysis of enhancer hubs were partly based on previously described methods[21,22], and are described in Supplementary Note 13**.**

T2D-FG variant enrichments in regulatory annotations. Variant Set Enrichment (VSE)[23] was used to compute the enrichment of T2D and FG-associated variants in regulatory annotations, using lead SNPs from 109 loci (Supplementary Table 9), and is described in Supplementary Note 14.

GWAS meta-analysis of insulin secretion. 7,807 individuals from four population studies were included in these analyses: the Inter99 study (ClinicalTrials ID-no: NCT00289237) (n = 5,305)[24], the Health2008 cohort (n = 605)[25], the 1936 Birth Cohort (n = 709)[26] and the ADDITION-Pro cohort (n = 1,188)[27]. All study participants gave informed consent and studies were approved by the appropriate Ethical Committees in accordance with scientific principles of the Helsinki Declaration II.

In all cohorts, glucose-stimulated insulin secretion was evaluated by measurement of plasma glucose and serum insulin at 0, 30 and 120 minutes during a 75 g oral glucose tolerance test (OGTT). We calculated Insulinogenic index = (s-insulin at 30 minutes [pmol/l] - fasting s-insulin [pmol/l]) / p-glucose at 30 minutes (mmol/l). Individuals with known diabetes were excluded.

Two sample sets (Inter99 and Health2008) were genotyped by Illumina OmniExpress array and others by Illumina CoreExome array. Genotypes were called by Illumina GenCall algorithm. Genotype data were filtered for variants with call rate <98% and Hardy-Weinberg equilibrium $P < 10^{-5}$. Samples were excluded if they were ethnic outliers, had mismatch between genetic and phenotypic sex or had a call rate <95%.

Genotype data were imputed to the Haplotype Reference Consortium (HRC) reference panel v1.1[28] at the Michigan Imputation Server using Minimac3 after phasing genotypes into haplotypes with Eagle2[29]. Post-imputation SNP filtering included exclusion of variants MAF < 0.01 or info score < 0.70. In each cohort, association analysis was performed by applying a linear regression model including age and sex as covariates via SNPTEST [30]. The phenotype was rank-normalized within each cohort before analysis. A fixed-effects meta-analysis implemented in the R package *meta* [31] was finally performed.

Heritability estimates (see also Supplementary Notes 16 and 17). To estimate the polygenic contribution of different genomic annotations to GWAS-based heritability of T2D and related traits we applied the stratified LD Score regression method[32,33]. The method leverages the relationship between LD structure and association test statistics to estimate the average per-SNP contribution to heritability (τ_c coefficient) of functional genomic categories. We used a baseline panel of 53 baseline genomic annotations[32,33], and interrogated broad range of islet regulatory annotations including enhancer hubs, as well as control annotation sets such as Central Nervous System functional annotations, random non-open chromatin regions, and pseudo-enhancer hubs. We provide the per-SNP heritability τ_c coefficient for each regulatory annotation. To facilitate comparisons across traits and annotations, we normalized the τ_c estimates by dividing them by the LD Score heritability for each phenotype, and multiplied by 10^7. To correct for multiple testing, we generated τ_c *q*-values (FDR-adjusted *P* values calculated from the Z-scores of the τ_c coefficients) with the *qvalue* R package over 17 functional categories and 6 traits. FDR significance threshold was set at 0.05.

Polygenic risk scores (PRS)(see also Supplementary Note 18). We created PRS based on T2D GWAS summary statistics from 70kfort2d[34] (*base dataset*). UK Biobank individuals [35] were used as the target datasets, which comprised training and testing datasets.
To select markers for PRS we first considered all genetic markers that were used as input for phasing and genotype imputation by UK Biobank, and filtered for variants with MAF ≥ 5% and imputation

quality score > 0.8. We then reconciled the base and target datasets by looking at the variant overlap between summary statistics and the imputed UK Biobank data, discarding variants showing allele inconsistency between both datasets. We also removed those located in the MHC region, resulting in a final collection of 5,352,737 variants.

We excluded UK Biobank individuals with: (i) excess of relatives (showing > 10 putative third-degree relatives, as provided by UK Biobank), (ii) greater than third-degree of relatedness (from each pair of related individuals we excluded the subject with the highest missing rate for a set of high-quality markers, as provided by UK Biobank), (iii) no gender information, (iv) ICD10 codes E10 (insulin-dependent diabetes mellitus), E13 (other specified diabetes mellitus) and E14 (unspecified diabetes mellitus), (iv) no body mass index (BMI) information. T2D cases were defined by the E11 ICD10 code.

The sample size of UK Biobank qualifying individuals was 377,981 controls and 15,764 cases, which was divided in training and testing datasets. For the training dataset, we included only control subjects with age at recruitment ≥ 55 years and no family history of diabetes mellitus, yielding a final training dataset sample size of 6,305 T2D cases and 73,922 controls. The remaining 236,236 individuals were used as a test dataset, and were not filtered by age or family history. PRS models were calculated from abovementioned base and training datasets using the PRsice software[36] with default settings and clumping parameters *(--clump-r2 0.6 --clump-p 0.01*). We included 11 covariates in the analysis: the 7 principal components provided by UK Biobank investigators as well as BMI, age at recruitment, batch information, and sex.

We generated PRS models based on the following common genetic variants: (a) the entire genome-wide set shared by the training and testing dataset (total of 5,352,737 variants; 1,152 qualifying variants in the model), (b) variants overlapping hub pcHi-C baits and enhancers (total variants = 86,158; 179 qualifying variants in the model), (c) variants overlapping islet open chromatin regions, excluding islet hub baits and enhancers and those in LD ($r^2 > 0.1$) with islet hub index variants (total variants = 269,342; 160 qualifying variants in the model), (b) the remaining genome,

excluding variants overlapping islet hub regions or other islet open chromatin regions or those in LD with islet hub index variants (total variants = 4,913,005; 355 qualifying variants in the model).

To enable comparisons of PRS effects in stratified subgroups, we created regions with similar genomic space and distribution as hubs (pseudo-enhancer hubs). Pseudo-enhancer hubs were generated essentially as for LD score regression analysis, except that they resembled hubs used for PRS, in that they contained all enhancers and baits of hubs. We created 100 sets of ~1,000 pseudo-enhancer hubs by shuffling hub pcHi-C baits and their assigned enhancer fragments across randomly selected size-matched TADs, excluding those in TADs with real hubs, or if they crossed TAD boundaries. We then built PRS models using variants overlapping these pseudo-baits and pseudo-enhancers (average of 265 qualifying variants per pseudo-hub PRS model).

To assess PRS, we first stratified the entire UK Biobank test dataset (n = 236,236) in 40 bins, each one containing 2.5% of individuals ranked by the PRS score. To enable assessment of PRS for T2D stratified by BMI and age of diagnosis, all measures of T2D frequency were performed exclusively with the 6,127 T2D cases with known age of diagnosis, and diagnosed after 20 years of age, and all 226,777 controls, which were censored at enrolment to UK Biobank. We calculated either T2D frequency ratios in top vs. bottom bin, or the odds ratio for T2D in individuals with highest PRS scores (top 2.5% bin) vs. remaining individuals in the same age and BMI categories using a logistic regression model adjusted for the first seven principal components of ancestry, sex, age, BMI and batch information. We expressed values as z-scores relative to the distribution of 100 sets of pseudo-hub PRS to enable comparisons of hub scores in the different stratified subgroups.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data visualisation

Data from this study can be visualized in the following browsers: Islet regulome browser (*http://isletregulome.org/isletregulome*)[37], CHiCP browser (https://www.chicp.org)[38] and WashU Epigenome browser using this session link:

http://epigenomegateway.wustl.edu/browser/?genome=hg19&session=62hGf7nfcS&statusId=140947077

Data availability

Raw sequence reads from pcHi-C, RNA-seq, ChIP-seq, ATAC-seq and 4C-seq are available from EGA (*https://www.ebi.ac.uk/ega)*, under accession number EGAS00001002917. Processed data files for islet pcHi-C interactions, islet regulome annotations, enhancer-promoter assignments, hub coordinates and components and 3D model videos are provided as supplementary data. The robust set of ATAC-Seq peaks, consistent set of Mediator, cohesin, H3K27ac and H3K4me3 peaks, list of islet super-enhancers defined using ROSE algorithm, islet regulome, ChromHMM segmentation model, list of islet TAD-like domains, PATs and the list of high-confidence pcHiC interactions are provided as Supplementary Data Sets and also deposited at https://www.crg.eu/en/programmes-groups/ferrer-lab#datasets.

Code Availability

Custom code in this manuscript is available upon request.

References for Methods

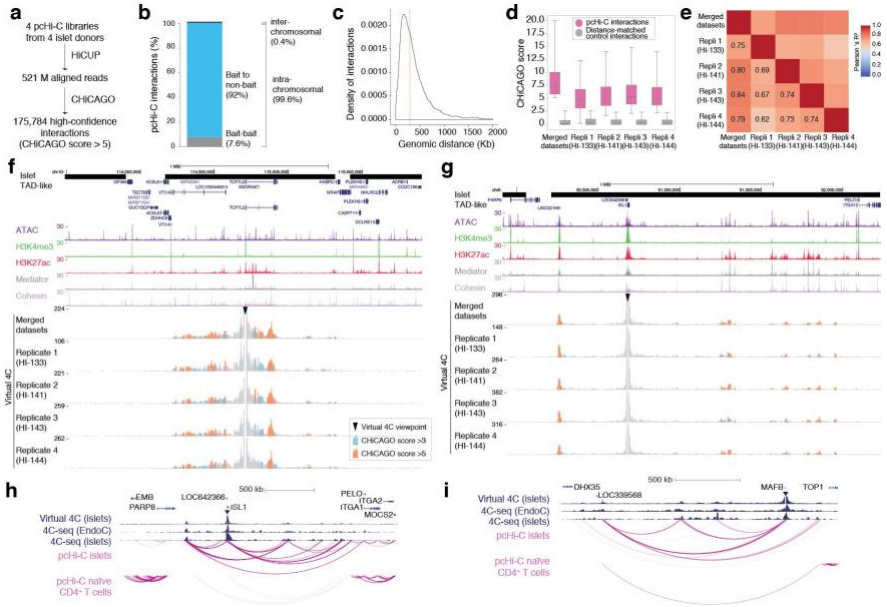
1. Melzi, R. *et al.* Role of CCL2/MCP-1 in islet transplantation. *Cell Transplant* **19**, 1031-46 (2010).
2. Kerr-Conte, J. *et al.* Upgrading pretransplant human islet culture technology requires human serum combined with media renewal. *Transplantation* **89**, 1154-60 (2010).
3. Bucher, P. *et al.* Assessment of a novel two-component enzyme preparation for human islet isolation and transplantation. *Transplantation* **79**, 91-7 (2005).
4. Ricordi, C., Lacy, P.E., Finke, E.H., Olack, B.J. & Scharp, D.W. Automated method for isolation of human pancreatic islets. *Diabetes* **37**, 413-20 (1988).
5. Nagano, T. *et al.* Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol* **16**, 175 (2015).
6. Javierre, B.M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384 e19 (2016).

7. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310 (2015).
8. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol* **17**, 127 (2016).
9. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* **46**, 136-143 (2014).
10. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-8 (2013).
11. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**(2011).
12. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
13. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
14. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
15. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
16. Kharchenko, P.V., Tolstorukov, M.Y. & Park, P.J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26**, 1351-9 (2008).
17. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
18. Leisch, F. A toolbox for K-centroids cluster analysis. *Comput. Stat. Data Anal.* **51**, 526-544 (2006).
19. Benazra, M. *et al.* A human beta cell line with drug inducible excision of immortalizing transgenes. *Mol Metab* **4**, 916-25 (2015).
20. Kuleshov, M.V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-7 (2016).
21. Bau, D. & Marti-Renom, M.A. Genome structure determination via 3C-based data integration by the Integrative Modeling Platform. *Methods* **58**, 300-6 (2012).

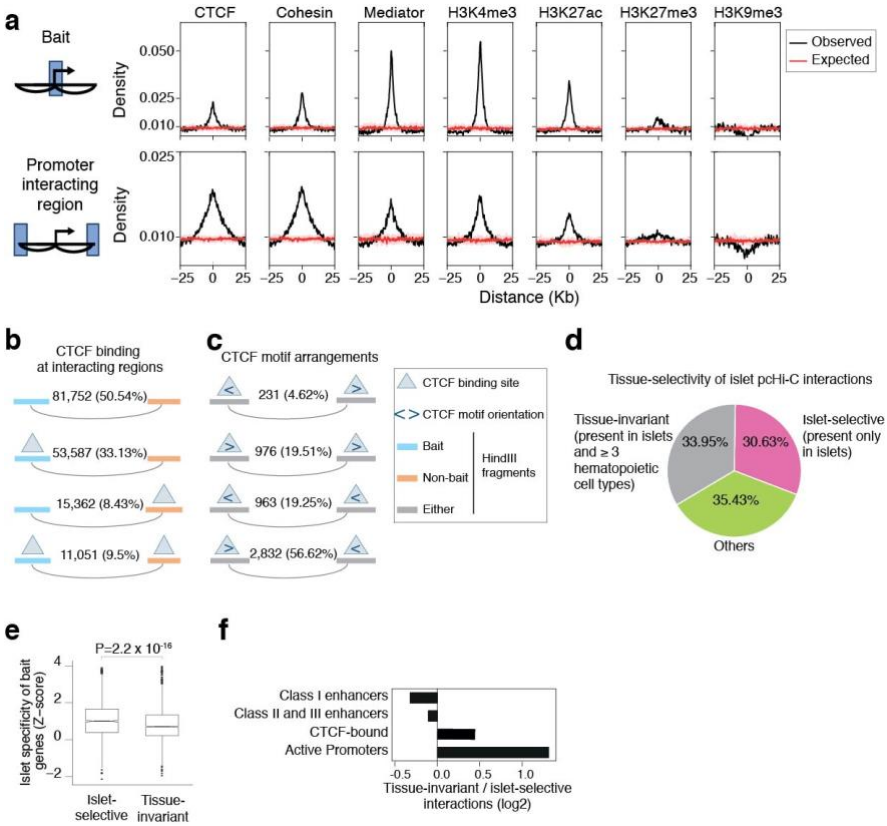
22. Di Stefano, M., Paulsen, J., Lien, T.G., Hovig, E. & Micheletti, C. Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. *Sci Rep* **6**, 35985 (2016).
23. Ahmed, M. *et al.* Variant Set Enrichment: an R package to identify disease-associated functional genomic regions. *BioData Min* **10**, 9 (2017).
24. Gjesing, A.P. *et al.* Genetic and phenotypic correlations between surrogate measures of insulin release obtained from OGTT data. *Diabetologia* **58**, 1006-12 (2015).
25. Thuesen, B.H. *et al.* Cohort Profile: the Health2006 cohort, research centre for prevention and health. *Int J Epidemiol* **43**, 568-75 (2014).
26. Drivsholm, T., Ibsen, H., Schroll, M., Davidsen, M. & Borch-Johnsen, K. Increasing prevalence of diabetes mellitus and impaired glucose tolerance among 60-year-old Danes. *Diabet Med* **18**, 126-32 (2001).
27. Johansen, N.B. *et al.* Protocol for ADDITION-PRO: a longitudinal cohort study of the cardiovascular experience of individuals at high risk for diabetes recruited from Danish primary care. *BMC Public Health* **12**, 1078 (2012).
28. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).
29. Loh, P.R., Palamara, P.F. & Price, A.L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet* **48**, 811-6 (2016).
30. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).
31. Schwarzer, G. meta: An R package for meta-analysis. *R News* **7**, 40-45 (2007).
32. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
33. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).
34. Bonas-Guarch, S. *et al.* Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. *Nat Commun* **9**, 321 (2018).

35. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
36. Euesden, J., Lewis, C.M. & O'Reilly, P.F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466-8 (2015).
37. Mularoni, L., Ramos-Rodriguez, M. & Pasquali, L. The Pancreatic Islet Regulome Browser. *Front Genet* **8**, 13 (2017).
38. Schofield, E.C. *et al.* CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. *Bioinformatics* **32**, 2511-3 (2016).

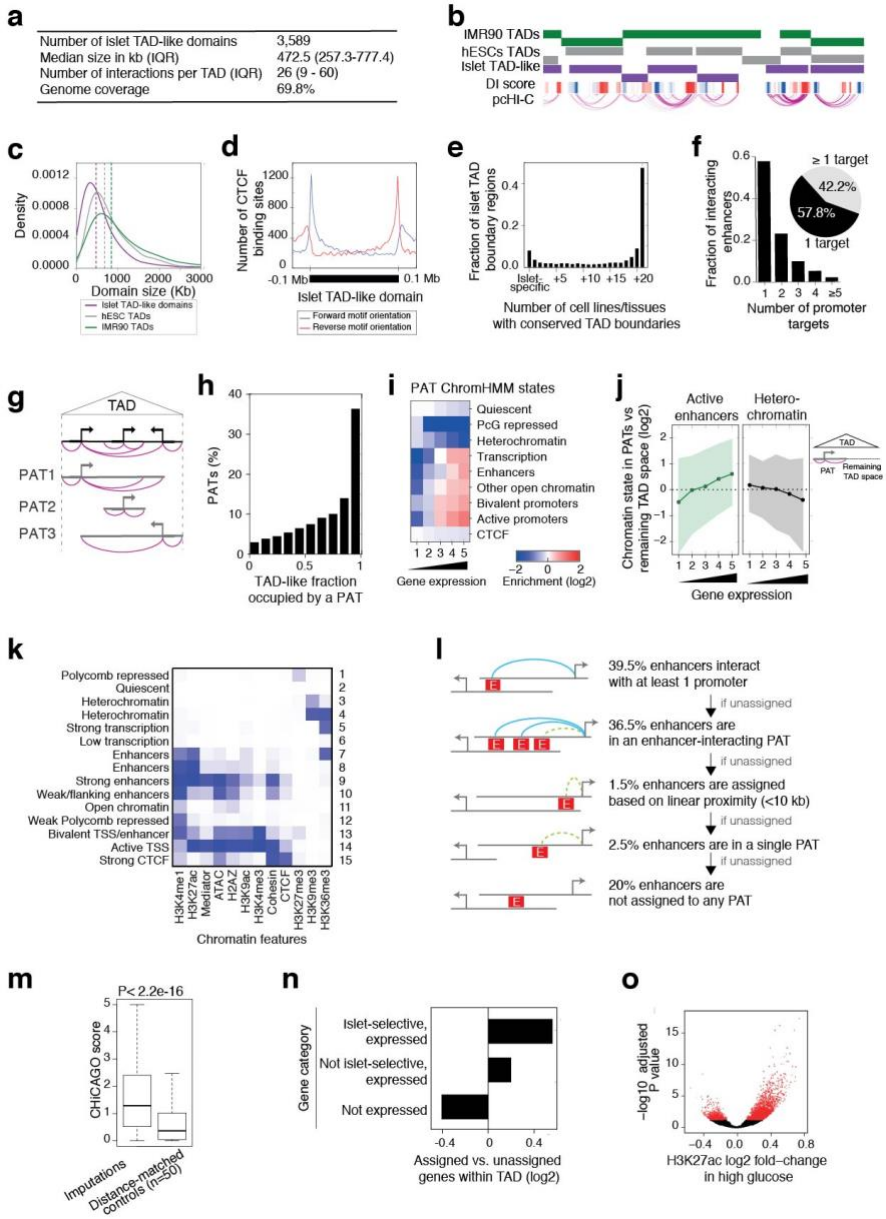
Supplementary Figures



Supplementary Figure 1. pcHi-C in human pancreatic islets.. a, Schematic representation of the pcHi-C analysis workflow. **b,** Relative frequency of high-confidence interactions between baits and interacting regions. **c,** Distances from bait to interacting regions for high-confidence interactions. The dashed line represents the median distance. **d,** CHiCAGO score distribution of high-confidence interactions in merged pcHi-C data (n=175,784) and individual islet samples , and in distance-matched interactions. Boxplots show IQR, and whiskers show 5th and 95th percentiles. **e,** Pairwise Pearson correlation values of CHiCAGO scores between individual islet samples and merged dataset. **f-g**. Epigenomic maps and virtual 4C profiles in merged and individual human islet samples in *TCF7L2* and *ISL1*. **h,i**. pcHi-C recapitulates interactions identified by 4C-seq in human islets and the human β cell line EndoC-βH1 at *ISL1* and *MAFB* loci. The top track depicts a virtual 4C representation of human islet pcHi-C data in both promoters. High-confidence interactions from 4 pooled human islet samples and naïve CD4+ T cells are shown below. Inverted triangles depict viewpoints.

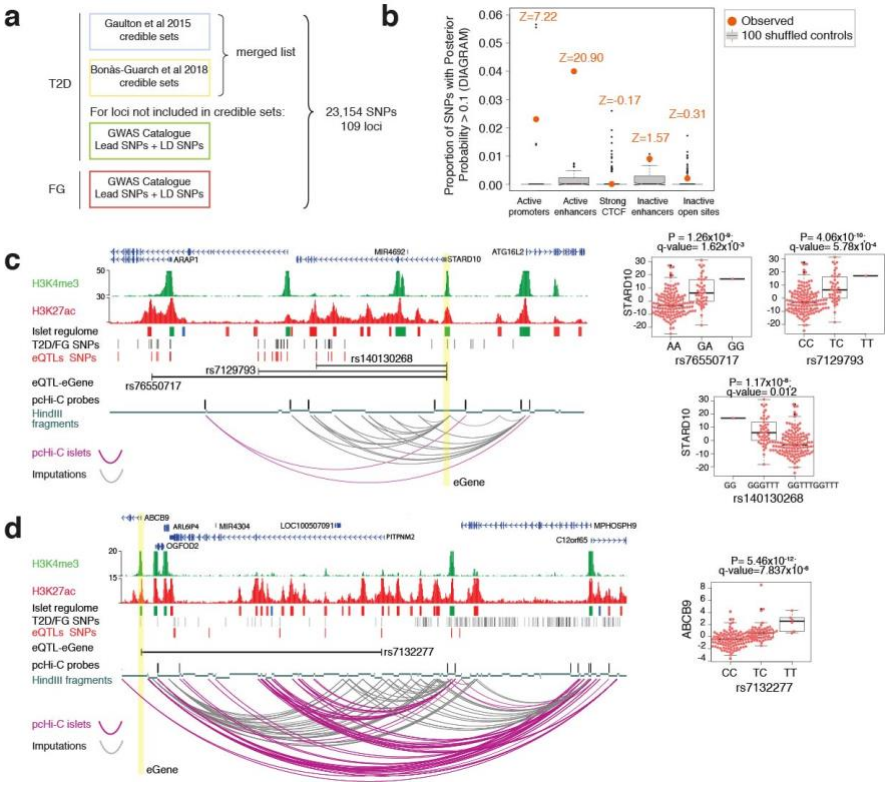


Supplementary Figure 2. pcHi-C and chromatin landscape of human islets. **a,** Binding patterns for indicated epitopes in ± 25 Kb regions centered on interacting pcHi-C baits (top), and promoter-interacting regions (bottom). Expected occupancy profiles after randomizing 10 times the positions of indicated signals are represented with a red line, and IQR are shown as a shade. **b,** Relative frequency of CTCF binding sites in baits and non-bait interacting regions. Nearly 50% of interactions are associated with CTCF binding in at least one of the interacting regions. **c,** CTCF-binding motif orientation at CTCF-bound interacting regions. 56.62% of 9,657 interactions are convergent, consistent with expectations. **d,** Tissue-selectivity of islet pcHi-C interactions relative to identically processed pcHi-C from erythroblasts, macrophages, naïve $CD4^+$ T cells and total B lymphocytes. **e,** Genes located in baits with islet-selective interactions show increased gene expression islet-specificity scores vs. genes with tissue-invariant interactions. The islet-specificity Z score was calculated with a gene expression distribution from 18 human tissues. P value was calculated with Wilcoxon's two-sided signed ranked test. Boxplot represents IQRs. **f,** Ratio of tissue-invariant to islet-selective interactions overlapping major open chromatin classes, normalized by the total number of tissue-invariant and islet-selective interactions. All categories showed significant differences with interactions in the remaining genome (Fisher's $P < 0.01$).

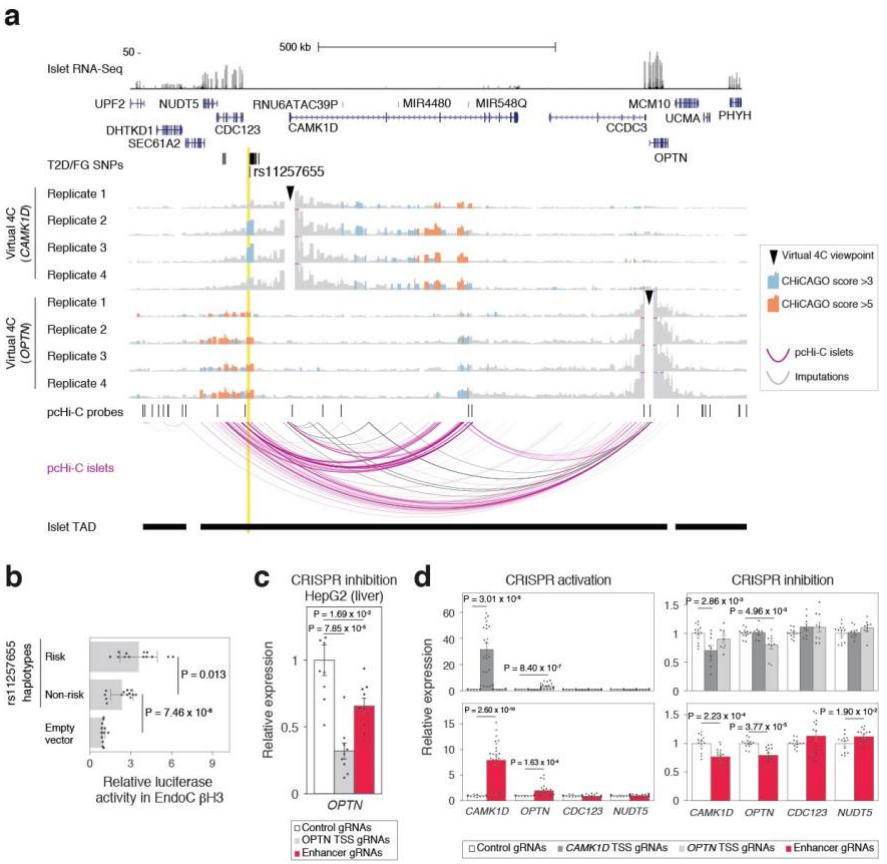


Supplementary Figure 3. Definition of TAD-like domains, PATs, and enhancer-gene assignments. **a,** Features of islet TAD-like domains. **b,** Representative example of human islet TAD-like domains (chr 11:1132582-4719948, hg19). Negative and positive directionality index (DI) scores are represented in blue and red, respectively. ESC and IMR90 TADs generated with Hi-C are shown for reference. **c,** Size of TAD-like domains in human islets and Hi-C TADs from ESC and IMR90 cells. **d,** TAD-like domains display known features of TADs, such as enrichment of CTCF binding and convergent CTCF motif orientation in borders. **e,** Tissue-selectivity of islet TAD-like boundary regions was estimated by comparison

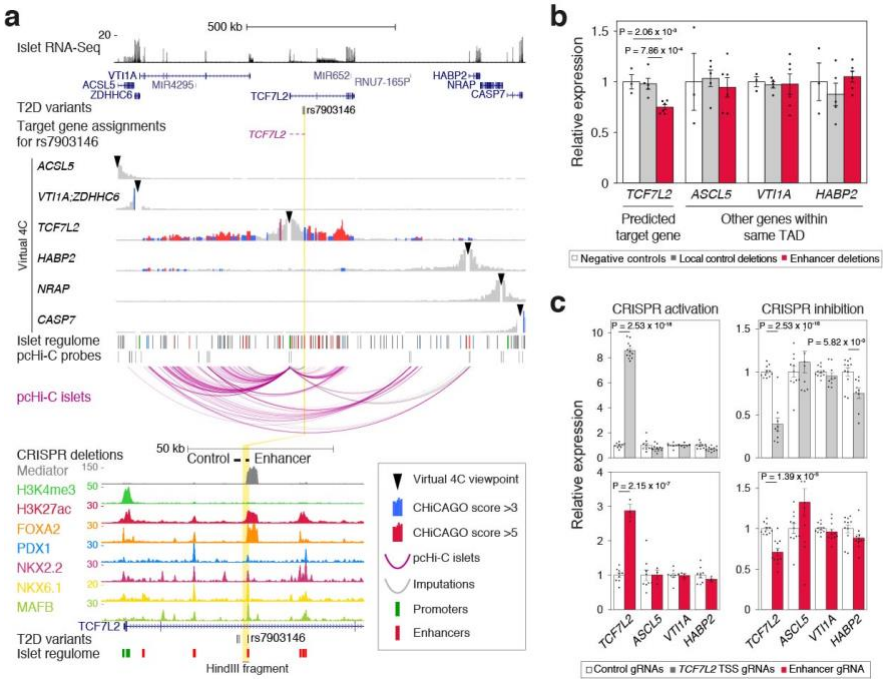
with TADs defined by Hi-C in 21 tissues. **f,** Enhancers frequently interact with more than one gene. Fraction of enhancers showing high-confidence (CHiCAGO > 5) interactions to 1-5+ promoter "baits" in the same TAD. **g,** Schematic of promoter-associated three-dimensional spaces (PATs), defined as the genomic space that spans high-confidence interactions originating from one bait. **h**, Fraction of islet TAD-like spaces occupied by each PAT. **i**. ChromHMM state enrichments in PATs were consistent with the expression level of their associated genes. The heatmap shows ChromHMM state median log2 fold-enrichments in PATs over their genomic distributions, in 5 bins based on bait gene expression levels in human islets. **j**. Active islet enhancer or H3K9me3-enriched ChromHMM states in PATs were enriched over the remaining TAD-like space in accordance with islet expression of PAT genes. Only PATs at least 25% smaller than their TAD were used (n=7,085). Median enrichments (circles) and IQR (shade) are shown. **k**. Emission probabilities of the 15 ChromHMM states for all islet chromatin features used to create the model. **l**. Sequential steps used to impute the assignment of islet enhancers to target genes. **m**. CHiCAGO scores for imputed enhancer-promoter pairs vs. distance-matched controls (n=50 sets). P value is from Wilcoxon's two-sided signed rank test. Boxplot represents IQRs. **n**. Genes assigned to enhancers were enriched in islet-specific genes, as compared with unassigned control genes from the same islet TAD-like structure (Chi-square $P = 6 \times 10^{-08}$). **o**. Islet exposure to 4 mM vs. 11 mM glucose causes widespread induction of H3K27 acetylation in islet enhancers. Dots represent H3K27ac-enriched regions, and are red if Benjamini-Hochberg adjusted $P \leq 0.05$.



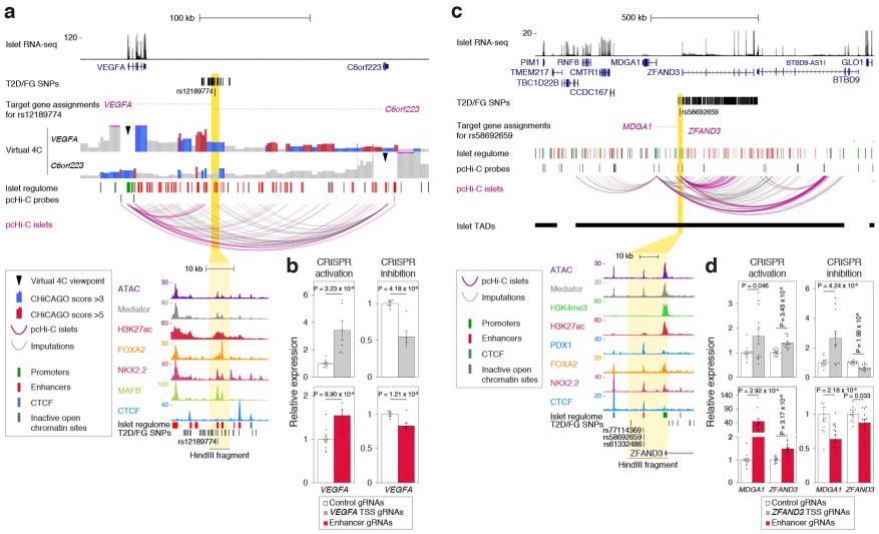
Supplementary Figure 4. eQTLs support the identification of unexpected T2D target genes. **a,** T2D and FG-associated variants used to examine gene targets (see Supplementary Table 3). **b,** Proportion of DIAGRAM credible set SNPs with high posterior probability (PP > 0.1) mapping to islet regulome elements within intervals containing credible sets. Note the enrichment in active enhancers and promoters vs.100 sets of elements shuffled within the genomic spaces that contain credible sets, shown as grey IQR boxplot distributions and outliers as black dots. Z-scores represent deviations from the mean of the shuffled distribution. **c-d,** Selected examples of loci with T2D-risk variants with gene targets supported by both significant eQTLs and pcHi-C, showing enhancer-gene assignments through pcHiC high-confidence interactions (from pooled data, in magenta) and imputations (grey). Enhancer eQTL-eGene pairs are represented as horizontal black lines. A vertical yellow stripe highlights the eGene promoter. Concordant gene targets include **c,** *STARD10* **d,** *ABCB9*. pcHiC interactions are represented as arcs connecting HindIII fragments. Boxplots shows first and third quartiles as boxes and 1.5 x IQR as whiskers of gene expression for different genotypes, shown as PEER residuals, along with P and adjusted P (q) values from eQTL meta-analysis. Red dots represent individual PEER residual values of gene expression for 183 samples across different genotypes. For additional eQTL findings see Supplementary Table 2.



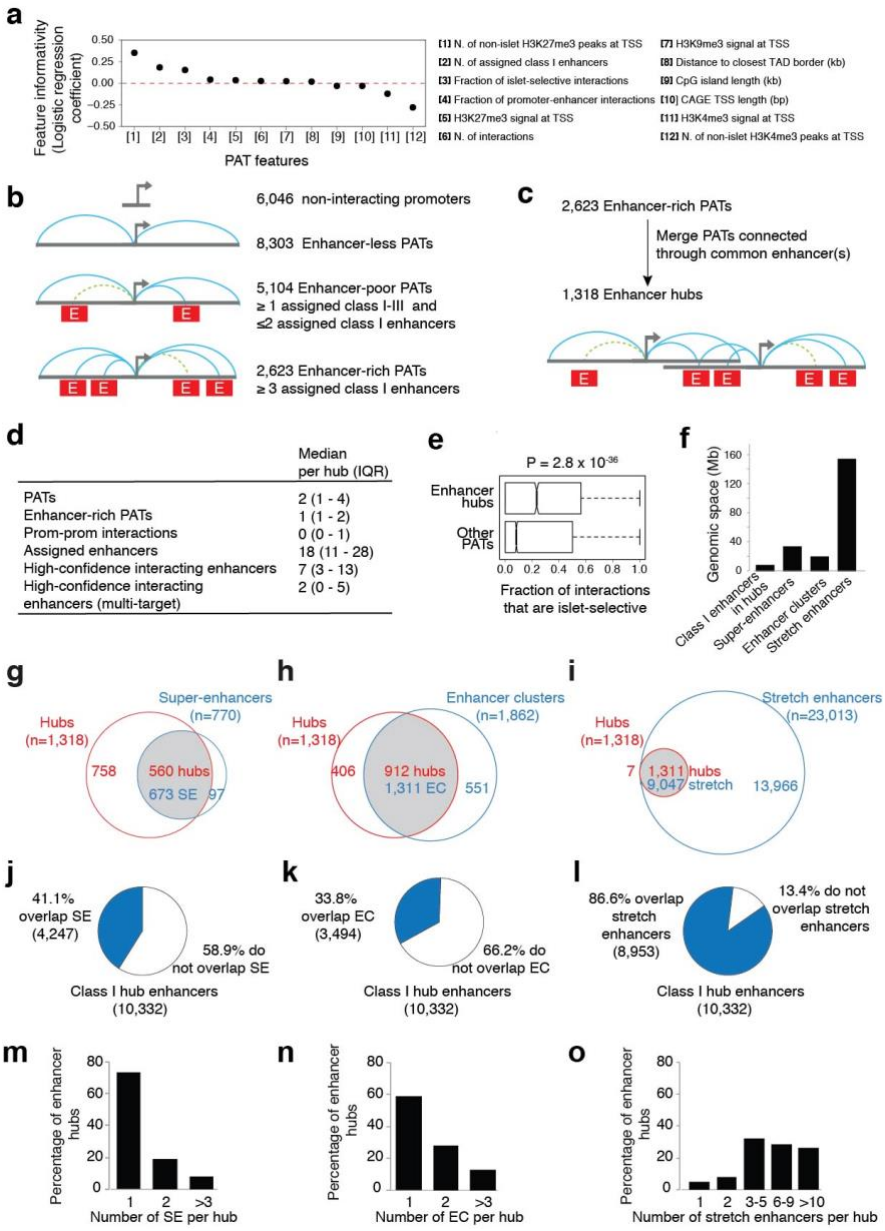
Supplementary Figure 5. **Functional perturbations of *CAMK1D* and *OPTN*. a,** Long-range interactions of the enhancer carrying rs11257655 are replicated in individual human islet pcHi-C samples. Note how interactions between this enhancer and *OPTN* are detected with high confidence (ChICAGO >5) in each pcHi-C replicate. **b,** Luciferase assay in the human β cell line EndoC-βH3 shows allele-dependent activity for the rs11257655-enhancer. Data are means ± s.d. (n=3 independent experiments, with 3-6 independent transfections). Statistical significance: two-tailed Student's *t*-test. **c,d**. Analysis of *OPTN* and *CAMK1D* mRNA after **c,** CRISPRi of the rs11257655-enhancer in HepG2 and **d,** CRISPRi or CRISPRa in EndoC-βH3 cells. Bars show average values of 3-4 gRNAs targeting either the rs11257655 enhancer, or the transcriptional start sites. Data are presented as means ± s.e.m. (enhancer activation: 4 gRNAs n=6; inhibition: 4 gRNAs n=3). Statistical significance: two-tailed Student's *t*-test.



Supplementary Figure 6. **Functional perturbations of *TCF7L2*. a,** Virtual 4C representations from pooled human islet samples centered on all genes in this locus show that the region containing rs7903146 connects with *TCF7L2* through moderate-confidence interactions and an imputed assignment, without evidence for interactions with other genes. The HindIII fragment that contains the enhancer with rs7903146 is highlighted in yellow. The bottom panel reveals that this enhancer shows unusually high occupancy by Mediator and islet-enriched transcription factors in islet chromatin. **b,** RNA analysis in EndoC-βH3 cells after deletion of either the rs7903146-enhancer or a control region in the same locus. Deletions were tested with 2 different gRNA pairs, n=3 experiments. Statistical significance was determined using two-tailed Student's *t*-test. Only active genes in the locus were tested. **c,** RNA analysis in EndoC-βH3 cells after CRISPRa or CRISPRi of the rs7903146-enhancer. Statistical significance was determined using two-tailed Student's *t*-test (activation: 1 gRNA, n=3 experiments; inhibition: 3 gRNAs n=3 experiments).

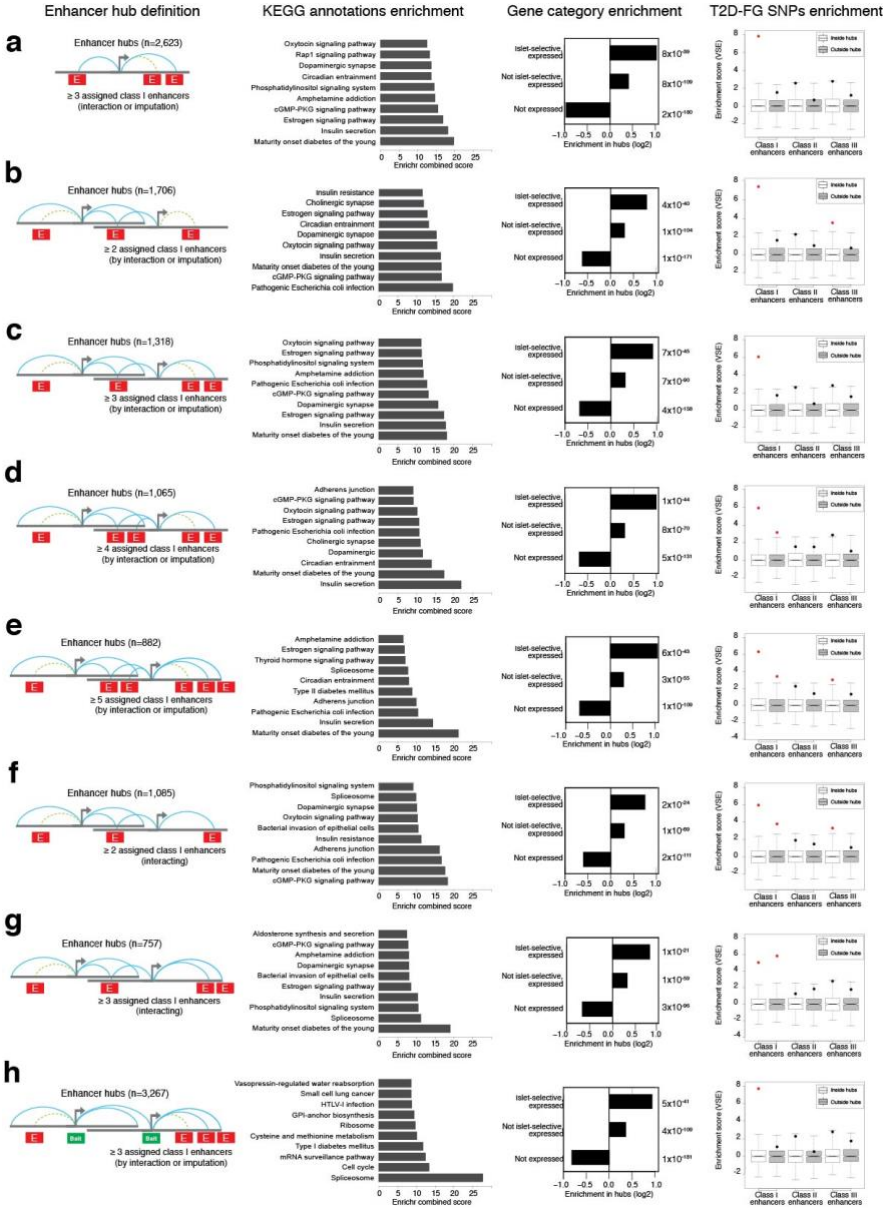


Supplementary Figure 7. **Functional perturbations of *VEGFA* and *ZFAND3*. a,c**. T2D variant-target gene assignments in *VEGFA* and *ZFAND3* loci. pcHi-C and virtual 4C representations are from pooled samples. **b,d**. *VEGFA* or *MDGA1* and *ZFAND3* mRNAs in EndoC-βH3 cells after CRISPRa or CRISPRi of T2D-associated enhancers. *C6orf223* was not detectable by qPCR. Note that we did not examine all potential targets near *VEGFA* (see other imputed genes in Supplementary Table 3). Data are presented as means ± s.e.m. (*VEGFA* enhancer CRISPRa: 3 guides n=3 experiments; *VEGFA* enhancer CRISPRi: 4 guides n=2 experiments; *ZFAND3-MDGA1* enhancer: 4 guides n=3 experiments). Statistical significance was determined using two-tailed Student's *t*-test.



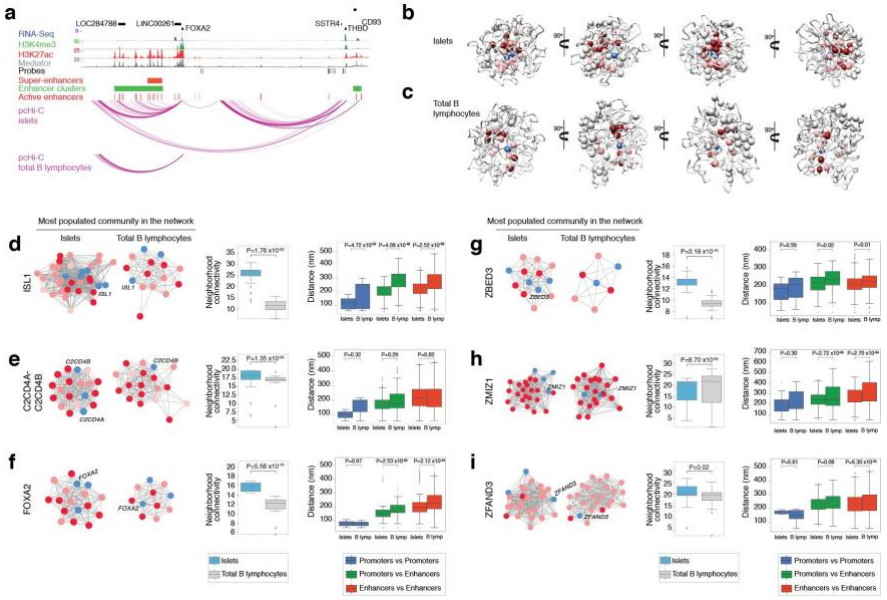
Supplementary Figure 8. **Tissue-specific enhancer hubs.** a, Multiple logistic regression analysis was used to identify PAT features that predict islet-expressed genes with islet-selective vs. non islet-selective expression. Islet-selective expression was examined as a surrogate endpoint because it is a property of many (though not all) genes important for islet cell identity. The PAT feature with the highest logistic regression coefficient was the number of non-islet tissues with promoter H3K27me3-enrichment. This feature was considered as almost synonymous with islet-specific islet expression. The next highest coefficient was the number assigned

class I enhancers in the PAT. Further analysis showed that ≥3 assigned class I enhancers in a PAT optimized the prediction of islet-selective expression (Supplementary Figure 9). b, Classification of PATs based on assigned enhancers revealed 2,623 enhancer-rich PATs (≥3 assigned class I enhancers). Enhancers are shown as red boxes. Turquoise and dashed green lines are high-confidence interactions and imputed assignments, respectively. c, Enhancer hubs were defined as enhancer-rich PATs, which were merged with other PATs connected through at least one common enhancer-associated high-confidence interaction. d, Descriptive characteristics of enhancer hubs in human islets. Multi-target enhancers show high confidence interactions with two or more promoter-containing baits. e, Enhancer hubs are enriched in islet-selective interactions relative to non-hub PATs that had at least 1 high-confidence interaction. Boxes are IQR, notches are 95% CI of the median and P values are from Wilcoxon's two-sided signed rank test. f, Linear genomic space occupied by class I enhancers in three-dimensional enhancer hubs compared with the space occupied by super-enhancers (SEs) calculated with the ROSE algorithm, all enhancers from linear enhancer clusters (ECs), and stretch enhancers. g-i. Venn diagrams depicting how often hub enhancers overlap with other human islet enhancer domains: g, SEs, h, highly-bound (top two TF occupancy quartiles) ECs, and i, stretch enhancers. j-l. Islet enhancer hubs often contain enhancers that do not form part of SEs or ECs. Charts show the fraction of hub class I enhancers that overlapped SEs, ECs or stretch-enhancers. Note that the genomic space occupied by stretch enhancers is an order of magnitude greater than hubs (panel g). m-o. Islet enhancer hubs very frequently contain multiple SEs, ECs or stretch enhancers.

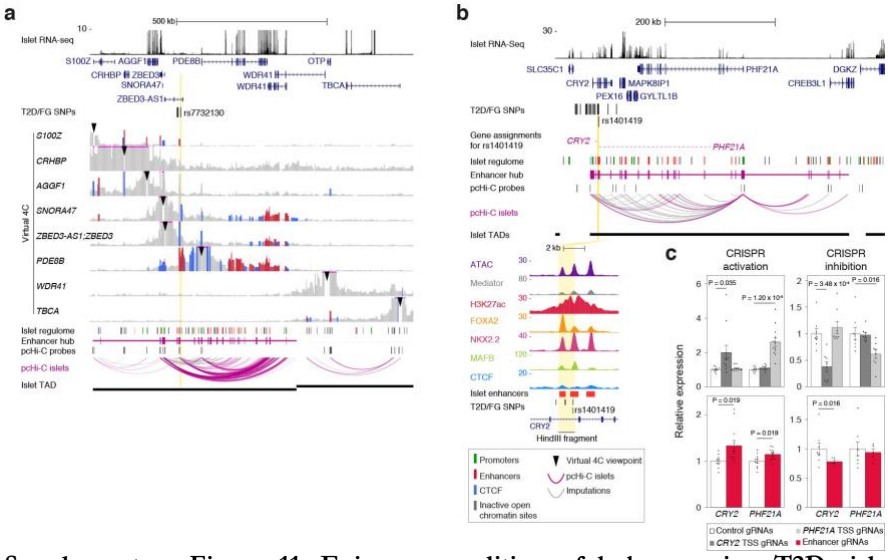


Supplementary Figure 9. **Alternative definitions of enhancer hubs**. We considered alternative definitions of hubs as follows: **a,** enhancer-rich PATs with ≥3 class I enhancers, but without merging interconnected PATs, **b-e**, enhancer-rich PATs with ≥2-5 assigned class I enhancers, merged with PATs interconnected through high-confidence enhancer interactions, **f,g,** enhancer-rich PATs with ≥2 or ≥3 class I enhancers exclusively assigned through high-confidence interactions, and then merged to PATs interconnected through high-confidence enhancer interactions, **h,** enhancer-rich PATs with ≥3 assigned class I enhancers, merged to

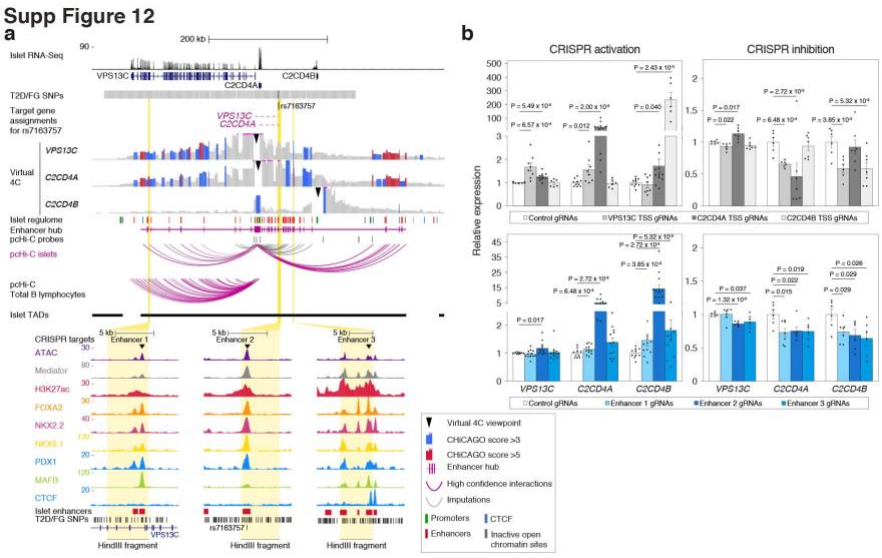
PATs interconnected through promoter-promoter (instead of enhancer-promoter) interactions. We found that canonical islet-cell functional annotations ranked highest only in definitions with ≥3 assigned class I enhancers. Hubs with ≥4-5 assigned class I enhancers (**d,e)**, as well as those defined exclusively with high-confidence interactions (**f,g)**, showed high ranking islet cell functional annotation enrichments, at the expense of reducing the number of hubs. Panels in the right show post-hoc VSE analysis of T2D/FG-associated SNPs (n=2,771; Supplementary Table 9). Consistent with the notion that the hub definitions in **d-g** were restrictive, they failed to show selective enrichment of T2D/FG-associated SNPs. Boxplots show null distributions based on 500 permutations of matched random haplotype blocks. Red dots indicate significant enrichment relative to the null distribution (Bonferroni–adjusted $P < 0.01$).



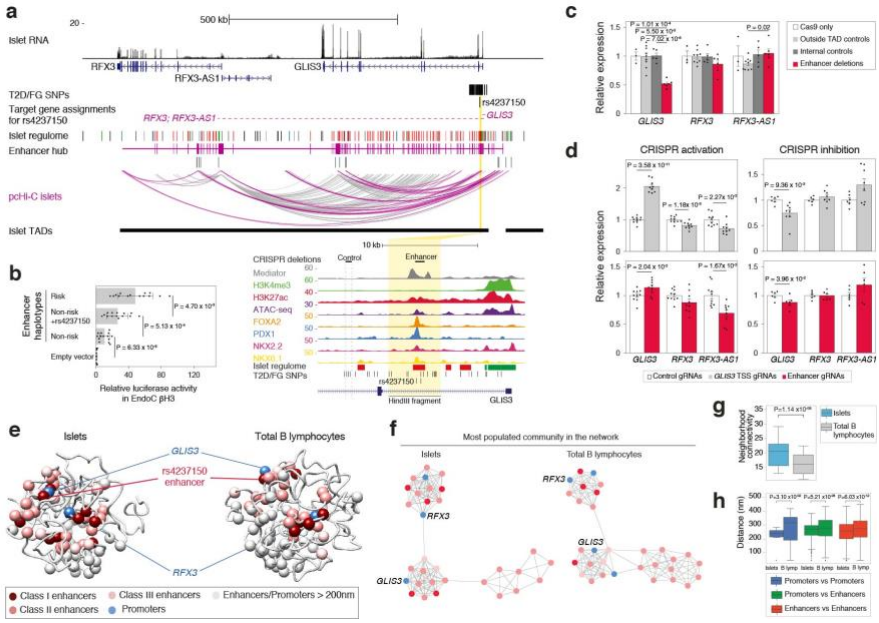
Supplementary Figure 10. **3D models of enhancer hubs**. **a,** The *FOXA2* locus forms a tissue-specific enhancer hub. Human islet epigenome maps and high-confidence pcHi-C interactions in islets and total B lymphocytes show that islet active enhancers, super-enhancers and enhancer clusters interact to form a single tissue-specific three-dimensional structure. **b-c**, 360º views of top-scoring 3D model of *ISL1* enhancer hub in human islets and total B lymphocytes. Class I, II and III enhancers within 200 nm of *ISL1* promoter are colored dark to light red, while promoters within 200 nm of *ISL1* (including *ISL1*) are colored blue. Islet enhancers and promoters are otherwise represented as white spheres. These models show that active islet regulatory elements interact in a common restricted space in islet nuclei. See also Supplementary Videos 1 and 2. **d-h,** Left panels show the most populated community of the promoter-enhancer interaction network in chosen hubs, as obtained via MCODE clustering, in human islets and total B lymphocytes. Network nodes are promoters (blue) and enhancers (dark to light red for enhancer classes I to III). Edges are mean distance values in the most populated 3D structure cluster. The central panel compares the neighborhood connectivity distribution of networks in both tissues. The right panel shows the 3D distances between hub promoters and enhancers in both tissues. All boxplots show IQRs and outliers as grey diamonds. The number of nodes analysed for each locus is shown in Supplementary Table 16. Statistical significance was computed using two-sided Kolmogorov-Smirnov test.



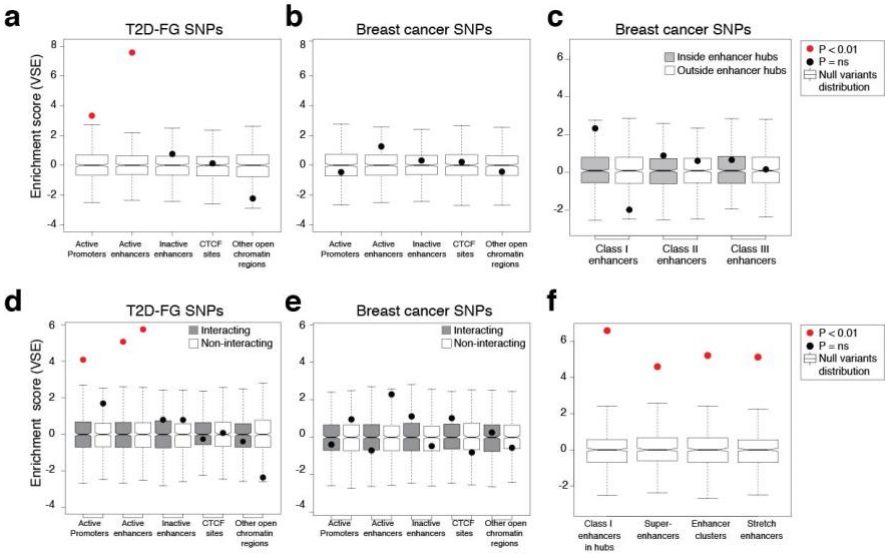
Supplementary Figure 11. **Epigenome editing of hubs carrying T2D risk noncoding variants**. **a,** pcHi-C and virtual 4C representations from pooled human islet samples in the *ZBED3* locus for all promoters with active transcripts in the region. **b,** Islet pcHi-C assigns *CRY2* and *PHF21A* as gene targets of an enhancer containing a FG-associated variant (vertical yellow stripe). **c,** Analysis of *CRY2* and *PHF21A* mRNA after CRISPRa or CRISPRi of their transcriptional start sites or of the islet enhancer bearing the FG-associated variant rs1401419 in EndoC-βH3 cells. Data are presented as means ± s.e.m. (enhancer CRISPRa: 4 gRNAs n=3; CRISPRi: 2 gRNAs n=2). Statistical significance was determined using two-tailed Student's *t*-test.



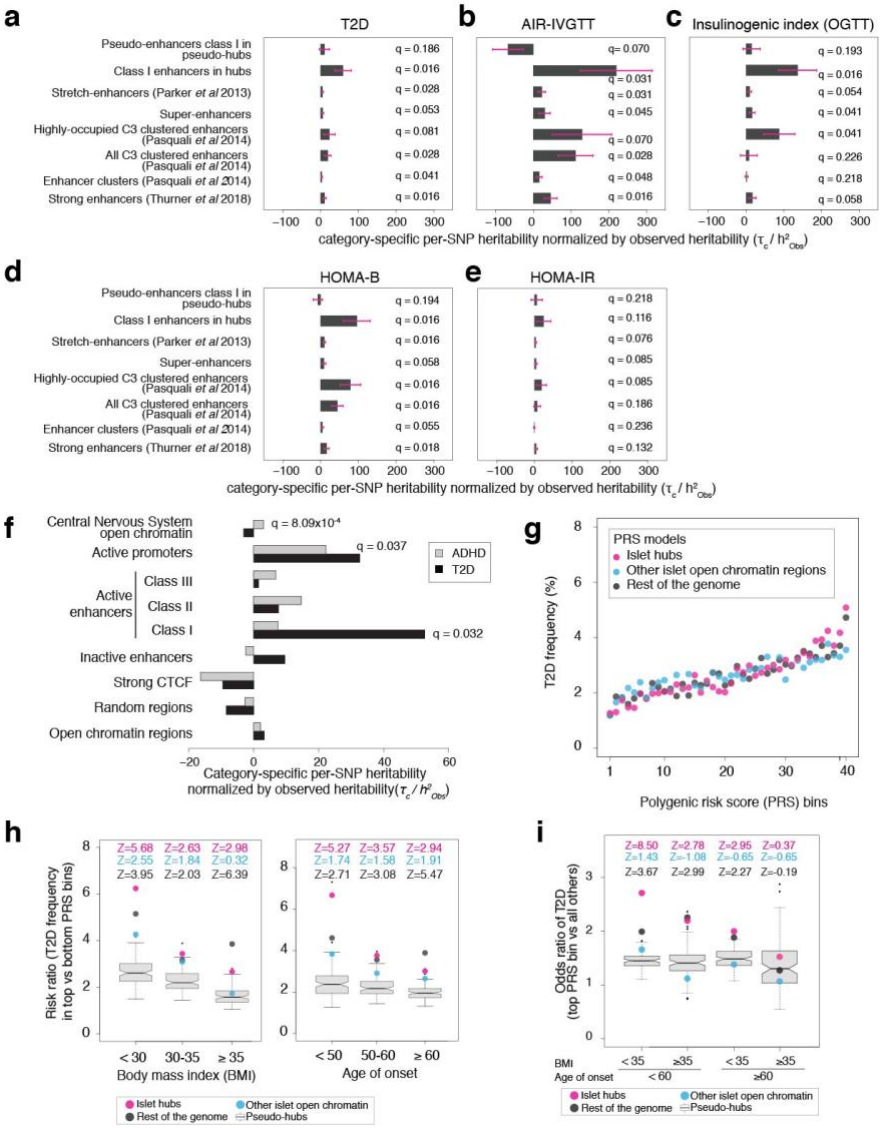
Supplementary Figure 12. **Epigenome editing of the *C2CD4A/B* hub**. **a,** Islet pcHi-C assigns *C2CD4A* and *C2CD4B* as gene targets of three enhancers containing T2D-associated variants (vertical yellow stripes) in the *C2CD4A/B* locus. pcHi-C and virtual 4C representations are from pooled human islet samples.
b, Analysis of *VPS13C, C2CD4A* and *C2CD4B* mRNA after CRISPRa or CRISPRi targeting of their transcriptional start sites or of three islet enhancers bearing T2D-FG variants in EndoC-βH3 cells. Data are presented as means ± s.e.m. (CRISPRa: 4 gRNAs n=3 experiments; CRISPRi: 4 gRNAs n=2 experiments). Statistical significance was determined using two-tailed Student's *t*-test.



Supplementary Figure 13. **Epigenome editing of the *GLIS3* hub**. **a,** Islet pcHi-C virtual 4C representations from pooled samples, showing the T1D/T2D-associated locus *GLIS3*. The inset shows the enhancer bearing rs4237150. **b,** Luciferase assays in EndoC-βH3 cells show haplotype-dependent activity of the rs4237150-enhancer. Data are means ± s.d. (n=3 independent experiments with 4-6 independent transfections). Statistical significance: two-tailed Student's *t*-test. **c,** Analysis of *GLIS3, RFX3* and *RFX3-AS1* mRNA upon deletion of rs4237150-enhancer or control regions. Data are presented as means ± s.e.m. (2 pairs of gRNAs per target region, n=3 experiments each). Statistical significance: two-tailed Student's t-test. **d,** Analysis of predicted target gene transcripts after CRISPRa or CRISPRi targeting of the *GLIS3* transcriptional start site or the rs4237150-enhancer in EndoC-βH3 cells. Data are means ± s.e.m. (enhancer CRISPRa: 3 gRNAs n=3 experiments; CRISPRi: 3 gRNAs n=2 experiments). Statistical significance: two-tailed Student's *t*-test. **e,** Top-scoring *GLIS3* hub model from the most populated cluster of the ensemble in human islets and total B lymphocytes. Enhancers and promoters within 200 nm *GLIS3* or *RFX3* promoters are colored in red and blue, respectively, or as white spheres if located further. **f,** Most populated community of the promoter-enhancer interaction network obtained via MCODE clustering of this locus in human islets and total B lymphocytes. Nodes represent promoters (blue) and enhancers (dark to light red for enhancer classes I to III). Edges are mean distances in most populated 3D cluster. Although *GLIS3 and RFX3* are connected in a common hub, the networks suggest that they form part of separable sub-communities. **g,** Neighborhood connectivity distribution between the islet and total B lymphocytes networks. **h,** 3D distance distribution between enhancers and promoters in *GLIS3* hub. Boxplots show IQRs. Statistical significance was computed using two-sample Kolmogorov-Smirnov two-sided test as described in Supplementary Figure 10. See also Supplementary Table 16.



Supplementary Figure 14. **T2D-associated variants are enriched in interacting regions and hub class I enhancers**. **a,b**, VSE enrichment analysis of T2D and FG (n=2,771) and breast cancer (n=3,048) variants in islet active regulatory elements (see **Supplementary Data Set 1**). Box plots show null distributions based on 500 permutations of matched random haplotype blocks. Each dot denotes VSE enrichment of disease-associated variants in each genomic feature. The red dot indicates significant enrichment relative to the null distribution (Bonferroni-adjusted $P < 0.01$). **c,** Breast cancer-associated variants show no enrichment in islet enhancer sub-classes. **d-e**, VSE enrichment analysis of T2D and FG and breast cancer SNPs in chromatin regions with high-confident pcHi-C interactions in islets. **f,** VSE enrichment analysis of T2D and FG-associated variants in indicated enhancer categories. All boxplots show IQRs.



Supplementary Figure 15. Class I enhancers in hubs contribute to heritability of beta cell-related traits. **a-e**, Per-SNP heritability estimates of variants in eight islet enhancer domain subtypes calculated using summary statistics data from: **a,** T2D (12,931 cases, and 57,196 controls); **b,** acute insulin release (AIR)-in vivo glucose tolerance test (IVGTT, up to 5,567 individuals); **c,** insulinogenic index (OGTT, 7,807 individuals); **d,** HOMA-B; and **e,** HOMA-IR (up to ~80,000 individuals). Bars show category specific per-SNP heritability coefficients (τ_c) divided by the LD score heritability (h^2) score observed for each trait. All normalized τ_c coefficients were multiplied by 10^7 and shown with s.e.m. τ_c coefficients were estimated using stratified LD score regression, controlling for 53 functional annotation categories included in the baseline model. **f,** Per-SNP T2D and Attention-

Deficit/Hyperactivity Disorder (ADHD, up to 55,374 individuals) heritability estimates in islet regulatory elements and Central Nervous System (CNS) annotations. τ_c coefficients, normalizations by h^2 and representations are as explained in panels a-e. **g,** Impact of polygenic risk scores (PRS) on T2D frequency. T2D frequency (y-axis) was calculated in 40 bins, each one representing 2.5% of individuals in the UK Biobank test set. PRS values were calculated with common genetic variants in islet hub enhancers and baits (pink dots), other islet open chromatin regions (light blue dots) and in the rest of genome (black dots). **h.** T2D risk ratios stratified by BMI (left) and age of onset of T2D (right). Controls were censored at the age of recruitment. Boxplots show IQR of the risk ratio from 100 sets of pseudo-hubs PRS, and with whiskers 1.5 x IQR. Color dots as in **g**. **h**, T2D risk stratified by BMI and age of onset of T2D. Odds ratios (OR) for T2D were calculated for 2.5% individuals with the highest PRS vs. all other individuals via adjusted logistic regression. Boxplots show IQR of the risk ratio from 100 sets of pseudo-hubs PRS, and with whiskers 1.5 x IQR. For all panels, Z-scores define standard deviations relative to average values from pseudo-hub PRS. See also Supplementary Figure 15 and Supplementary Table 17.

Supplementary Tables and rest of Supplementary information available online:

https://www.nature.com/articles/s41588-019-0457-0#Sec33

DISCUSSION

3D reconstruction of genomic regions from sparse interaction data

In the work presented in Chapter 1 of this thesis, we have introduced a new integrative modelling protocol for the normalisation, 3D reconstruction, and analysis of data coming from sparse 3C-based experiments. Specifically, we have optimised our tool for pcHi-C experiments and tested its limitations with decreasing levels of captures in a synthetic dataset that can represent a generic capture experiment. Moreover, we have demonstrated its usability for the differential analysis of cell-type-specific chromatin architectural features, using as an example the β-globin locus.

As stated in the Introduction, 3C-based experiments rely on the usage of a diverse set of tools for normalisation (*Hu, Deng et al. 2012*, *Imakaev, Fudenberg et al. 2012*, *Servant, Varoquaux et al. 2015*, *Wingett, Ewels et al. 2015*, *Durand, Shamim et al. 2016*, *Vidal, le Dily et al. 2018*) and further analysis of the interaction data (*Lun and Smyth 2015*, *Djekidel, Chen et al. 2018*, *Ardakany, Ay et al. 2019*). In the case of pcHi-C, the novelty of the technique and the complexity associated to the sparseness of the data has resulted in a reduced number of available tools (*Cairns, Freire-Pritchett et al. 2016*, *Mifsud, Martincorena et al. 2017*, *Anil, Spalinskas et al. 2018*, *Ben Zouari, Molitor et al. 2019*, *Cairns, Orchard et al. 2019*). Most of these tools focus on detecting significant interactions from the experimental data, or on the comparative analysis between datasets. Conversely, the integrative modelling tool we have designed in this work allows the user to assess the significance of the measured distance between two selected loci of interest and to perform differential analysis between cell-types (and stages, see annex 1). Additionally, it contextualises the interaction data into a 3D space, facilitating its interpretation and further analysis based on spatial enrichment of selected features, and most importantly, recovering the organisation of the full loci despite of the data sparseness.

Firstly, we tested our procedure by comparing models reconstructed from sparse (pcHi-C) and dense (Hi-C) datasets. This comparison showed that the reconstructed sparse models similarly recovered the organisation from the dense ones when using both virtual pcHi-C datasets inferred from the dense Hi-C and real pcHi-C datasets. These results further indicated that the biases coming from the capture protocol had efficiently been reduced by our method. Additionally, and most importantly, the sparse models efficiently recovered most of the structure from the dense models, thus allowing the analysis of particles that had not been interrogated in the experimental assay.

Next, we used synthetic toy genome models (*Trussart, Serra et al. 2015*) to measure the minimum amount of restraints needed to reliably recover the architecture of a defined genomic region. Surprisingly, with just 2-3% of all possible interaction data from the matrix, we achieved a median correlation greater than 0.8 using ten random sets of capture distributions. In light of these results, we suggest that our integrative modelling protocol might also be useful for the 3D reconstruction of other sparse 3C-based datasets like HiChIP (*Mumbach, Rubin et al. 2016*), among others.

Finally, we tested the utility of our method by applying it to the β-globin locus, whose 3D organisation has been extensively studied before (*Schübeler, Francastel et al. 2000*, *Palstra, Tolhuis et al. 2003*, *Brown, Leach et al. 2006*, *Huang, Keller et al. 2017*, *Liu, Zhang et al. 2017*). We focused in the structural comparison of cord-blood Erythroblasts (cb-Ery), naïve $CD4^+$ T-cells (nCD4), and Monocytes (Mon), where the β-globin locus was active (cb-Ery) or inactive (nCD4 and Mon) in a cell-type-specific manner. In agreement with previous works (*Javierre, Burren et al. 2016*), analysis of the 3D topology of these cells showed different conformations associated with the activity stage of the β-globin locus. Interestingly, our models showed an enrichment of expression and active epigenetic marks around HBG2, the most expressed gene of the locus in cb-Ery. This functional signature was absent in the β-globin-inactive cell types (nCD4 and Mon), where the β-globin locus occupied a region depleted of expression and active chromatin marks.

We also show that this enrichment arises as a consequence of the gathering in space of the haemoglobin genes with loci located at long genomic distances (> 1 megabase) in cb-Ery, but not in nCD4 and Mon. Remarkably, our models show the formation of a 3D network that segregates the different cell-type-specific subsets of expressed genes in communities. This organisation is compatible with previous findings describing the gathering in space of transcribed genes as a general mechanism to organise gene transcription (*Jackson, Hassan et al. 1993*, *Osborne, Chakalova et al. 2004*, *Fraser and Bickmore 2007*, *Osborne, Chakalova et al. 2007*, *Baù, Sanyal et al. 2011*, *Sanyal, Bau et al. 2011*). Further analysis showed that these communities have cell-type-specific community stabilities (as defined by the co-occurrence score values of the communities within the model ensemble), with more stable communities in cb-Ery as compared with the more unstable ones of nCD4 and Mon. Thus, stability metrics of expressed gene communities might be important features for the identification of cell-type-specific 3D signatures. Additionally, we observed that in cb-Ery, but not in the gene communities of nCD4 and Mon, both gene communities and genes embodied inside of each community, overall arranged following an expression gradient, with the most expressed entities placed in the centre, and the least ones on the periphery. Based on these evidences, we hypothesize that the defined communities might represent cell-type-specific transcription factories (*Iborra, Pombo et al. 1996*, *Sutherland and Bickmore 2009*, *Baù, Sanyal et al. 2011*, *Sanyal, Bau et al. 2011*) or phase-separated foci (*Boija, Klein et al. 2018*, *Cho, Spille et al. 2018*, *Gurumurthy, Shen et al. 2019*). This would explain the gradient of transcription in terms of transcription machinery concentration in the core of the communities, with higher expression of the genes that are closer to it. Since this hierarchy of expression is not present in nCD4 and Mon, we suggest it is a cell-type-specific 3D signature characterising the β-globin region in cb-Ery.

Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes

In the work presented in Chapter 2 of this thesis, we apply our integrative modelling protocol for the normalisation, 3D reconstruction, and analysis of human pancreatic islet pcHi-C data in the context of enhancer-promoter 3D clusters relevant for the development of type 2 diabetes (T2D). Specifically, this work linked diabetes-associated enhancers with their target promoters, defining a list of ~1,300 3D enhancer hubs which are enriched in T2D associated signals and show glucose-dependent activity. The 3D enhancer hubs were also enriched in T2D risk variants, and further validation by genome editing of 8 selected loci showed their reliability to detect regulatory elements relevant for the development of T2D.

In this collaborative effort, we used a previous version of our normalisation approach (PRINT), which instead of obtaining a proportion of interaction between the interacting bins, weighted their value by the summation of the whole-genome interactions of the least interacting bin (see **Methods** in **Chapter 2**). We then used our integrative modelling protocol to reconstruct the 3D organisation of seven T2D-relevant hubs in pancreatic islets and B lymphocytes. The epigenetic profile of pancreatic islets was used to define the enhancer location coordinates inside of each of the hubs. These coordinates, together with the already known locations of the captured promoters were used in islets and B lymphocyte data-derived models to analyse the differential structural organisation of the hubs between both cells. Specifically, we measured the distances between enhancers and promoters to both build networks and calculate their neighbourhood connectivity, and to obtain and compare their distance distributions.

Further analysis of the models and the networks built from the pcHi-C datasets showed that islet-specific enhancers and their target promoters overall colocalised in a more constrained space in islets than in B lymphocytes, thus forming more connected enhancer-promoter networks, and highlighting the cell-type-specific

colocalization of multiple interspersed genomic regions to form defined 3D hubs inside of TADs.

CONCLUSIONS

From chapter 1, we can specifically conclude that:

1. We developed an integrative 3D modelling protocol to reconstruct the architecture of the chromatin from sparse 3C-based datasets.
2. We optimised this protocol for the normalisation, 3D reconstruction and differential analysis of pcHi-C datasets.
3. The method reconstructs highly similar structures overcoming most of the different experimental biases coming from Hi-C and pcHi-C.
4. The method retrieves reliable models with as low as 2-3% of all the possible interactions from the interaction matrix.
5. The method is accurate enough to recapitulate the known structural organisation of the β-globin locus and cell-type-specific arrangements associated with the level of expression of the involved loci.
6. We introduced innovative tools for the differential analysis of genomic 3D structures.

From chapter 2, we can specifically conclude that:

1. Our tool can be used to reconstruct the sub-TAD organisation of 3D enhancer hubs.
2. Distance data retrieved from the chromatin 3D models can be used to build regulatory elements networks.
3. The data subtracted from the models can be used to perform differential organisation analysis that help characterizing cell-type-specific conformations relevant for the development of Diabetes Type II.

ANEX 1

CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response

Grégoire Stik, Enrique Vidal, Mercedes Barrero, Sergi Cuartero, Maria Vila-Casadesús, Julen Mendieta-Esteban, Tian V. Tian, Jinmi Choi, Clara Berenguer, Amaya Abad, Beatrice Borsari, François le Dily, Patrick Cramer, Marc A. Marti-Renom, Ralph Stadhouders and Thomas Graf. **CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response**. Nature Genetics. 51, 1137–1148(2019)

CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response

Grégoire Stik[1*], Enrique Vidal[1], Mercedes Barrero[1], Sergi Cuartero[1,2], Maria Vila-Casadesús[1], Julen Mendieta-Esteban[3], Tian V. Tian[1,4], Jinmi Choi[5], Clara Berenguer[1,2], Amaya Abad[1], Beatrice Borsari[1], François le Dily[1], Patrick Cramer[5], Marc A. Marti-Renom[1,3,6], Ralph Stadhouders[7,8*] and Thomas Graf[1,9*]

[1] Centre for Genomic Regulation (CRG) and Institute of Science and Technology (BIST), Barcelona, Spain.
[2] Josep Carreras Leukaemia Research Institute (IJC), Barcelona, Spain
[3] CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.
[4] Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain
[5] Max Planck Institute for Biophysical Chemistry, Göttingen, Germany.
[6] ICREA, Barcelona, Spain.
[7] Department of Pulmonary Medicine, Erasmus MC, Rotterdam, the Netherlands.
[8] Department of Cell Biology, Erasmus MC, Rotterdam, the Netherlands.
[9] Universitat Pompeu Fabra (UPF), Barcelona, Spain.

*Corresponding authors: *gregoire.stik@crg.eu*; *r.stadhouders@erasmusmc.nl*; *thomas.graf@crg.eu*

SUMMARY

Three-dimensional (3D) organization of the genome is important for transcriptional regulation[1–7]. In mammals, CTCF and the cohesin complex create sub-megabase structures with elevated internal chromatin contact frequencies, called topologically associating domains (TADs)[8–12]. Although TADs can contribute to transcriptional regulation, ablation of TAD organization by disrupting CTCF or the cohesin complex causes modest gene expression changes[13–16]. In contrast, CTCF is required for cell cycle

regulation[17], embryonic development, and formation of various adult cell types[18]. To uncouple the role of CTCF in cell state transitions and cell proliferation we studied the effect of CTCF depletion during the conversion of human leukemic B cells into macrophages with minimal cell division. CTCF depletion disrupts TAD organization but not cell transdifferentiation. In contrast, CTCF depletion in induced macrophages impairs the full-blown upregulation of inflammatory genes after exposure to endotoxin. Our results demonstrate that CTCF-dependent genome topology is not strictly required for a functional cell fate conversion but facilitates a rapid and efficient response to an external stimulus.

MAIN TEXT

Lineage instructive transcription factors establish new cell identities by activating a novel gene expression program while silencing the old one. Whereas they largely achieve this through binding to promoters and enhancers, genome topology has recently emerged as a new player in gene regulation. Chromatin contact maps, obtained by chromosome conformation capture techniques such as Hi-C, revealed that chromatin can be separated at the megabase level into active ('A') and inactive ('B') compartments[19], themselves subdivided into TADs. Large deletions overlapping boundaries can cause a fusion of adjacent domains that can lead to developmental abnormalities[20]. In addition, inversion or deletion of individual CTCF binding sites can induce a loss of specific contacts or insulation from active chromatin[21–23]. Mechanistically, genomic insulation by TADs is thought to facilitate enhancer-promoter interactions while inhibiting cross-boundary communication between regulatory elements to prevent aberrant gene activation[24]. Hence, the importance of CTCF and TAD organization in facilitating transcriptional rewiring during cell state transitions – often accompanied by extensive cell division – remains controversial[25].

We have recently developed a system uniquely suitable to study the role of CTCF in cell state transitions, consisting of a B leukemia cell line (BLaER) that can be efficiently converted by exogenous CEBPA expression into functional induced macrophages (iMacs) with only one cell division on average (**Fig. 1a**; Supplementary Note 1)[26]. Using this system, we analyzed a time-series of

transdifferentiating cells for genome-wide changes in 3D genome organization (in-situ Hi-C), enhancer activity (ChIP-seq of histone modifications), chromatin accessibility (ATAC-seq) and gene expression (RNA-seq).

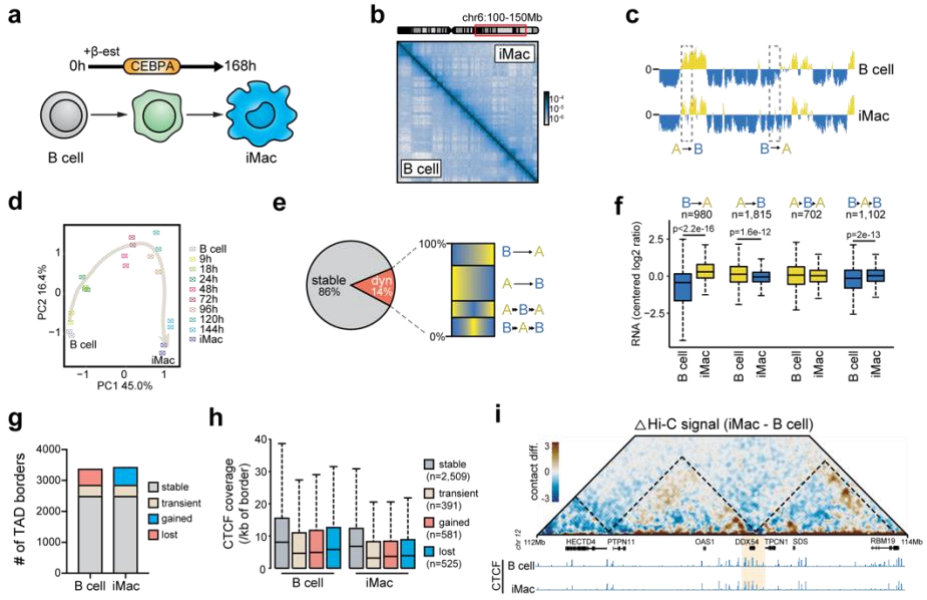


Fig. 1 | Transcription factor-driven transdifferentiation re-wires nuclear compartments and modulates TAD borders independently of CTCF binding. **a**, Schematic overview of the transdifferentiation system. CEBPA-ER in B cells (BLaER cell line) translocates to the nucleus after β-estradiol (β-est) treatment, activating the factor. A week after treatment the cells convert into induced macrophages ('iMac' stage). **b**, Representative in situ Hi-C contact maps (100-kb resolution) of a 50-Mb DNA region of B cells and iMacs. Color scale represents the normalized number of contacts per read. **c**, Transformation of the Hi-C map based on the PC1 values of a PCA on the Hi-C correlation matrix. PC1 values for A and B compartment are shown in yellow and blue, respectively; dotted rectangles highlight local compartment changes during transdifferentiation. **d**, PCA of PC1 compartment values (n = 28,749 bins), with grey arrow indicating transdifferentiation trajectory. **e**, Proportion of dynamic compartment bins (dyn.) including its distribution of different sub-categories. **f**, Integration of gene expression associated with dynamic compartment using RNA-seq (n represents the number of genes and *P* values are calculated using a two-sided Wilcoxon rank-sum test). **g**, Number of stable, transiently changed, gained or lost TAD borders in B cells and iMacs. **h**, CTCF-peak coverage at the different types of borders (n represents the number of borders in each category) in B cells and iMacs. **i**, Top: Differential contact map (iMac minus B cell signal) at the *DDX54* locus. Color scale represents differential contacts per 100,000 reads. Bottom: Snapshot of genome browser showing CTCF ChIP-seq signals at the locus. CTCF peaks at the newly created border are highlighted. All box

plots depict the first and third quartiles as the lower and upper bounds of the box, with a thicker band inside the box showing the median value and whiskers representing 1.5× the interquartile range.

We first determined genome segmentation into A and B compartments on the basis of the first eigenvector values of a principal component analysis (PCA) on the Hi-C correlation matrix ('PC1 values'). Overall, although most of the genome remained stable, around 14% of A or B compartments were dynamic during transdifferentiation, showing transcriptional changes correlating with the altered compartmentalization (**Fig. 1b-f, Extended Data Fig. 1a-d**; Supplementary Note 2). Next, we used chromosome-wide insulation potential[27] to identify between 3,100-3,300 TAD borders per time point (**Fig. 1g**). Boundaries were highly reproducible between biological replicates (Jaccard index > 0.99) and enriched in binding sites for CTCF (**Extended Data Fig. 1e**). Genome-wide insulation scores analyzed by PCA over time revealed progressive changes, reflecting a transdifferentiation trajectory (**Extended Data Fig. 1f**). While 70% of TAD borders were stable across all stages, 18% were lost or gained and 12% were transiently altered (**Fig. 1g**). CTCF binding was significantly more enriched at stable than at dynamic boundaries (**Fig. 1h**), as observed earlier[28]. Furthermore, while lost borders showed some CTCF occupancy in B cells that decreased in iMacs, gained borders were depleted for CTCF in both cell states (**Fig. 1h**), indicating CTCF-independent mechanisms driving local insulation. The dynamic rearrangement of TAD borders during transdifferentiation is illustrated by the *DDX54* locus (**Fig. 1i**), in which a new boundary appears in iMacs without apparent changes in CTCF binding. Furthermore, border gain or loss did not correlate with changes in local gene expression (**Extended Data Fig. 1g**), indicating that transcription is not a driver of the observed changes. However, whereas motif analysis at ATAC-seq peaks within stable borders indeed showed a strong enrichment for the CTCF motif, dynamic borders were enriched for PU.1 and EBF1 motifs (**Extended Data Fig. 1h**), raising the possibility that lineage-restricted transcription factors are involved in disrupting and/or establishing these borders.

To directly assess the importance of CTCF during CEBPA-induced transdifferentiation we devised an auxin-inducible degron

approach[29] (**Fig. 2a**; Supplementary Note 3). Addition of auxin to these cells triggered proteasome-dependent CTCF degradation, resulting in a loss of mCherry$^+$ cells and rapid CTCF depletion to levels undetectable by Western blot (**Fig. 2b-c**). Likewise, 80% of CTCF peaks were no longer detected after auxin treatment, and the enrichment level of persistent peaks was substantially reduced (**Extended Data Fig. 2a-b**), as previously described for mouse embryonic stem cells[14]. We next performed Hi-C on cells cultured in the presence of auxin or DMSO (as a control) at 24 and 168 hours post-induction (hpi) of transdifferentiation. Scaling of contact probabilities as a function of genomic separation did not change after CTCF depletion (**Extended Data Fig. 2c**). Analysis of chromosome-wide insulation potential in wildtype and CTCF-AID B cells showed that fusing the mAID tag to CTCF only had a negligible impact on TAD organization (**Extended Data Fig. 2d-e**). However, ~70% of TAD borders became undetectable and not visible in Hi-C contact maps after auxin treatment, both at 24 hpi and iMac stages (**Fig. 2d** and **Extended Data Fig. 2f**). Overall, insulation scores at borders detected in the control cells (DMSO) were dramatically reduced upon auxin treatment both at 24hpi and iMac stages (**Extended Data Fig. 2g**). Consequently, the ratio of contact enrichment inside TADs over outside was also strongly decreased (**Extended Data Fig. 2h**). Whereas stable borders exhibited a dramatic loss of insulation after CTCF depletion, dynamic borders showed essentially no change (**Fig. 2e**), in agreement with their low CTCF occupancy (**Fig. 1h**).

We next used ATAC-seq and H3K4me1/H3K27ac ChIP-seq to identify promoters and enhancers that are either activated, inactivated or remain stable during transdifferentiation (**Extended Data Fig. 2i-l**; see Methods). Transdifferentiation was accompanied by extensive chromatin state dynamics focused at enhancers, which were preferential targets of CEBPA binding (**Extended Data Fig. 2j-k**). We then interrogated how CTCF depletion affects intra-TAD enhancer-promoter (E-P) contacts at 0 h, 24 hpi and in iMacs. E-P interaction frequencies significantly decreased during inactivation, which was somewhat accelerated in auxin-treated cells at 24 hpi and iMac stages (**Fig. 2f**). Similarly, E-P interaction frequencies significantly increased during activation, while E-P interactions at stable regulatory elements were not affected by CTCF depletion (**Fig. 2f**). These data demonstrate that although auxin-treated

samples show a minor overall reduction of intra-TAD E-P contacts, E-P interaction dynamics accompanying transdifferentiation seem independent of CTCF.

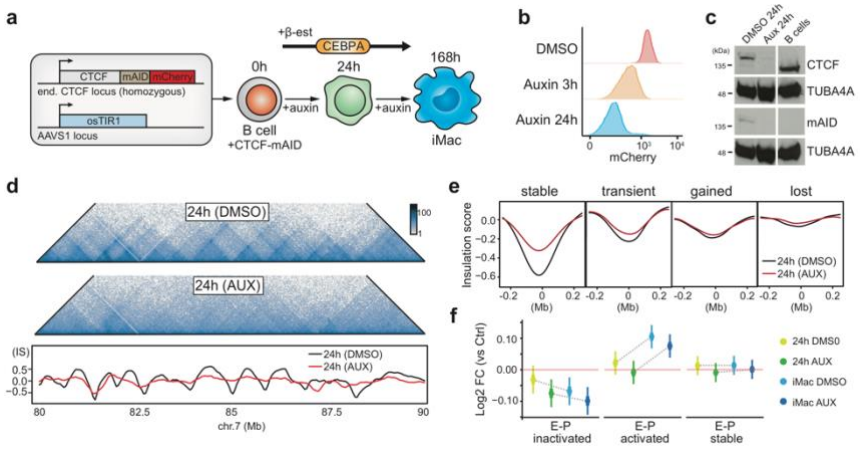


Fig. 2 | Auxin-mediated depletion of CTCF impairs chromatin insulation at stable but not dynamic TAD boundaries. a, Schematic representation of auxin-mediated CTCF degradation, showing the constructs used and the design of the experimental setup. **b,** Flow cytometry analysis showing decreased mCherry fluorescence intensity after auxin treatment as a proxy for CTCF levels. The experiment was repeated 3 times with similar results. **c**, Western blot showing loss of CTCF in CTCF-mAID B cells treated with auxin. Detection of TUBA4A was used as a loading control. The blots have been cropped from original blots available in Source Data. The experiment was repeated 3 times with similar results. **d,** Top: Representative Hi-C contact maps (20-kb resolution) of a 10-Mb region in chromosome 7 from transdifferentiating cells (24 hpi) treated with DMSO or auxin. Color scale represents the normalized number of contacts. Bottom: insulation score line graphs across the locus. **e**, Insulation scores at stable, transient, gained and lost borders of cells treated with DMSO or auxin. Areas shown are centered on boundary regions ± 250-kb. **f**, Changes of enhancer-promoter (E-P) intra-TAD contacts during transdifferentiation with DMSO (n = 2 biologically independent samples) or auxin (n = 2 biologically independent samples) in comparison to B cells (n = 1). Dots represent point estimates and bars (wide and narrow) indicate confidence intervals (50% and 95%, respectively) for the $\log_2$ fold changes. Estimations are computed using all 9 samples in a single linear mixed model.

To assess whether CTCF depletion and a loss of TAD organization impacts CEBPA-induced transdifferentiation, we monitored the expression of the B cell marker CD19 and the macrophage marker Mac-1 (CD11b) by flow cytometry at 0, 24, 48, 96 and 168 hpi. Surprisingly, CTCF-depleted cells converted into macrophage-like cells with even slightly accelerated kinetics at intermediate time points (**Fig. 3a** and **Extended Data Fig. 3a**), which was confirmed

using a different clone of CTCF-mAID B cells (**Extended Data Fig. 3b**). The iMacs obtained under conditions of CTCF depletion were phagocytic and activated inflammatory cytokine genes in response to endotoxin treatment (**Extended Data Fig. 3c-d**). Our findings show that CTCF depletion and widespread loss of TAD organization neither blocks nor delays transdifferentiation of B cells into functional macrophages.

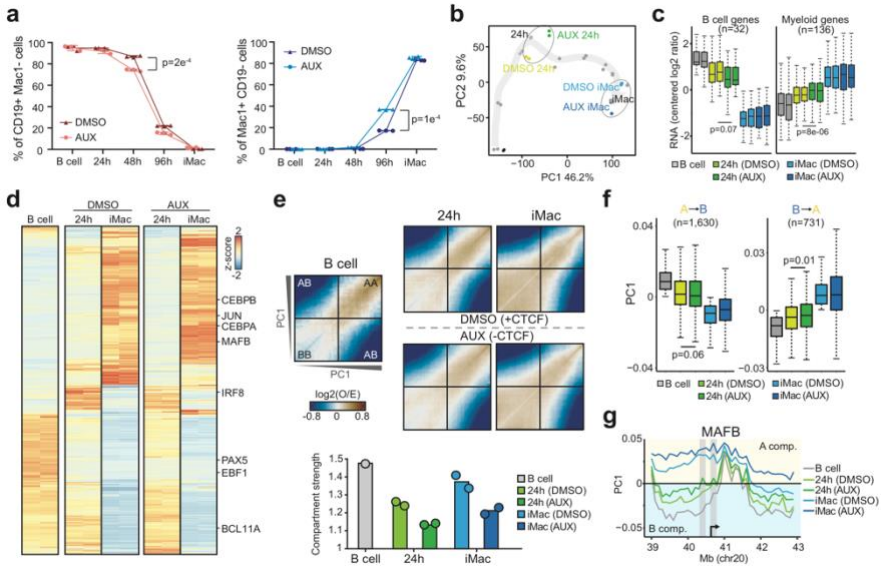


Fig. 3 | CTCF is dispensable for transcription factor-induced cell fate conversion. a, Flow cytometry analysis during transdifferentiation of cells treated with DMSO or auxin. Graphs show percentages of $CD19^{+}Mac1^{-}$ cells (left) or $CD19^{-}Mac1^{+}$ (right) cells (n = 3 biologically independent samples, error bars show standard deviation and *P* unpaired two-tailed *t* test). **b,** PCA analysis of transcriptome changes during transdifferentiation of CTCF-mAID B cells treated with DMSO or auxin (n = 23,680 genes). Grey points connected by an arrow represent non-tagged B cell transdifferentiation. Ellipses group 24 hpi and iMac stage samples. **c,** RNA expression of selected B cell (n = 32) and myeloid cell genes (n = 136) during transdifferentiation with DMSO and auxin for the biological 2 replicates (*P*, two-sided Wilcoxon rank-sum test). **d,** Heatmap of differentially expressed annotated transcripts (n = 2 biologically independent samples, FC > 2 and $P < 0.01$, two-tailed likelihood ratio test followed by FDR correction) in cells treated with DMSO or auxin. Myeloid and B cell regulator genes are indicated on the right. **e,** Top: saddle plot showing pairwise enrichment of the 20% top and bottom PC1 values from Hi-C contacts at 100-kb bins (see Methods). Lower part: compartmentalization strength scores derived from B cell (n = 1) and DMSO (n = 2) or auxin-treated cell (n = 2) biologically independent samples. The score corresponds to the ratio between same-compartment and different compartment contacts (diagonal corners over anti-diagonal corners in the saddle plots). **f,** Average

PC1 values of dynamic compartment bins (A to B n = 1,630 and B to A n = 731) in B cell (n = 1) and DMSO (n = 2) or auxin-treated cell (n = 2) biologically independent samples (*P*, two-sided Wilcoxon rank-sum test). **g,** Plot of PC1 values (100-kb bins) at the *MAFB* locus during transdifferentiation in the presence of DMSO or auxin. All box plots depict the first and third quartiles as the lower and upper bounds of the box, with a thicker band inside the box showing the median value and whiskers representing 1.5× the interquartile range.

We next analyzed how gene expression is affected upon CTCF depletion in cells at 24 hpi and at the iMac stage. A PCA of the entire transcriptome showed that CTCF depletion does not impair the overall rewiring of gene expression induced during cell fate conversion (**Fig. 3b**). Instead, auxin-treated cells were more advanced towards transdifferentiation at 24 hpi, which was further confirmed by analyzing gene expression dynamics of B cell and myeloid cell signature genes (**Fig. 3b-c**). These observations agree with previous findings suggesting that a partial knockdown of CTCF accelerates myeloid commitment of common myeloid precursor cells[30]. A heatmap of 8,595 annotated transcripts that changed significantly (fold change > 2, $P < 0.05$) during transdifferentiation highlighted the overall similarity between DMSO and auxin-treated samples (**Fig. 3d**). In fact, 76% of differentially expressed genes in control iMacs were similarly regulated under conditions of CTCF depletion (**Extended Data Fig. 3e**). This is illustrated for cell type-restricted transcription factors by activation of *CEBPB*, *JUN*, *CEBPA* and *MAFB*; by transient upregulation of *IRF8*; and by silencing of *PAX5*, *EBF1* and *BCL11A* in a similar fashion under both conditions (**Fig. 3d**). Although iMacs produced in the presence of auxin functionally resemble macrophages, they still show substantial differences in gene expression (~13% of expressed genes) compared to DMSO controls, mostly involving ubiquitous cellular processes like the cell cycle, GTPase signaling or ribosome biogenesis (**Extended Data Fig. 3f**).

The finding that CTCF depletion impacts TAD organization without substantially altering transdifferentiation capacity and kinetics prompted us to explore other features of 3D genome organization. Analyzing our Hi-C data for inter-TAD long-range E-P interactions (5 -10 Mb) revealed that their activation or inactivation is associated with the formation or dissolution of interacting clusters, respectively (**Extended Data Fig. 3g**). Remarkably, this occurred

independently of CTCF (**Extended Data Fig. 3g**), suggesting that the observed 3D clusters are linked to compartmentalization changes involving transcription factors bound to these regulatory regions. Further analyses revealed that 78% of the regions that switched during transdifferentiation from one compartment to another do so in both DMSO- and auxin-treated cells (**Extended Data Fig. 3h**), showing that CTCF is largely dispensable for these large-scale genome rearrangements. In line with a previous study[14], we observed ~10% reduction in compartment strength in auxin-treated samples (**Fig. 3e**), which could explain the slight acceleration of compartment transitions observed in auxin-treated cells at 24 hpi (**Fig. 3f**). An example is provided by the *MAFB* locus, a myeloid-expressed gene that is upregulated during transdifferentiation (**Extended Data Fig. 3i**), whereby the B-to-A switch was faster and more pronounced in auxin-treated cells than in DMSO controls (**Fig. 3g**), including at an enhancer region that becomes decorated with H3K27ac at 24 hpi (**Extended Fig. 3j**). In short, our Hi-C data revealed that although CTCF appears dispensable for genome compartmentalization, its depletion slightly decreased compartmentalization strength, which could facilitate compartmental rearrangements.

Previous reports indicated that CTCF plays a role in controlling macrophage gene expression[31] and that cohesin is required for an optimal inflammatory response of macrophages[32]. This raised the possibility of an involvement of genome topology in mounting an acute inflammatory response, as CTCF is known to stabilize the interaction between cohesin and chromatin[33,34]. Accordingly, aggregates of our Hi-C signals at previously described cohesin-bound loops[35] showed that these interactions disappeared after auxin treatment (**Extended Data Fig. 4a**). Using a public dataset of lipopolysaccharide (LPS) responsive genes[36] we found that both the enhancers and promoters of such genes are enriched for CTCF and that their promoters are closer to enhancers as compared to unresponsive genes (**Extended Data Fig. 4b-d**). Therefore, we tested the effect of CTCF depletion in iMacs exposed for 2 h to LPS (**Fig. 4a**), revealing a reduced induction of critical LPS-responsive genes such as *IL6, TNFA, CCL2* (**Fig. 4b**). Even more pronounced changes were observed at the level of secreted cytokines 8 h after LPS treatment (**Fig. 4c**). We next used RNA-seq to investigate the genome-wide effect of CTCF depletion on LPS-treated iMacs. Out

of 39,963 detected genes, 746 were found significantly upregulated ($P < 0.01$), although pathway enrichment analysis could not detect any significant associations (**Extended Data Fig. 4e-f**). Conversely, the 694 downregulated genes (among them *IL-6*, *TNFA* and *CCL2*) were strongly associated with pathways related to the inflammatory response to bacterial stimuli (**Extended Data Fig. 4e-f**). Although a sizeable fraction of differentially expressed genes were already altered in auxin-treated iMacs prior to LPS stimulation, the total number of affected genes doubled after LPS exposure (**Extended Data Fig. 4g**) and the observed upregulation upon LPS treatment was significantly blunted after CTCF depletion (**Extended Data Fig. 4h**). Of note, the expression of these genes was not significantly changed in CTCF-depleted iMacs prior to LPS treatment (**Extended Data Fig. 4i**) and most of the key transcription factors and receptors involved in the LPS response were unaffected after 24 h of auxin treatment (**Extended Data Fig. 4j**), suggesting a direct role for CTCF in fine-tuning the expression of inflammatory response genes. A similar proportion of promoters of upregulated or downregulated genes after CTCF depletion and LPS stimulation were bound by CTCF in iMacs (**Extended Data Fig. 4k**), indicating no dominant role for CTCF as a promoter-proximal repressor[37] in this context. A phagocytosis assay with DMSO- or auxin-treated iMacs showed that although CTCF-depleted cells were still functional, the number of engulfed beads per cell was reduced (**Extended Data Fig. 4l-m**), in line with the observed attenuation of the acute inflammatory response.

We next investigated whether CTCF-mediated 3D genome organization could underlie the apparent sensitivity of inflammatory response genes to CTCF depletion. Genes activated by LPS that were downregulated in CTCF-depleted iMacs were located closer to TAD borders and also more strongly insulated than random gene sets (**Extended Data Fig. 5a-b**), suggesting they could be extra-susceptible to deregulation by a loss of CTCF. To validate this and assess the impact of CTCF depletion on E-P interactions at key inflammatory response genes, we used our Hi-C data to conduct a 'virtual' 4C analysis of the *IL6* and *CCL2* loci, centered on their promoters. Active enhancers within these loci were identified by H3K27ac enrichment. At the *IL6* locus, CTCF depletion not only disrupted insulation from neighboring TADs but also decreased the frequency of *IL6* E-P interactions (**Fig. 4d** and **Extended Data Fig.**

5c). Interestingly, the neighboring gene *STEAPB1* located just upstream of the *IL6* TAD was found to be ectopically expressed upon CTCF depletion, likely resulting from aberrant contacts with *IL6* enhancers that are normally suppressed by the *IL6* TAD border (**Extended Data Fig. 5d-e**). To gain further insight into local chromatin conformation changes we generated 3D models of the *IL6* locus using Hi-C interaction data, transforming the interaction frequencies between genomic segments into spatial restraints[38]. This revealed that initially the *IL6* locus resides in a constrained space isolated from adjacent regions and that upon CTCF depletion the regions collapsed into less well-defined domains, separating the enhancers from their cognate target promoter (**Fig. 4e-f**). These models also confirm the decreased distance between *STEAPB1* and the *IL6* enhancers in the absence of CTCF (**Fig. 4f** and **Extended Data Fig. 5f**). Similar observations were made at the *CCL2* locus, where CTCF depletion also induced a loss of chromatin insulation and a decrease in E-P contacts (**Extended Data Fig. 5g-j**). These findings indicate that in macrophages, CTCF-mediated chromatin insulation and E-P interactions maintain acute inflammatory response genes in a primed configuration, permitting their rapid and robust activation in response to bacterial stimuli.

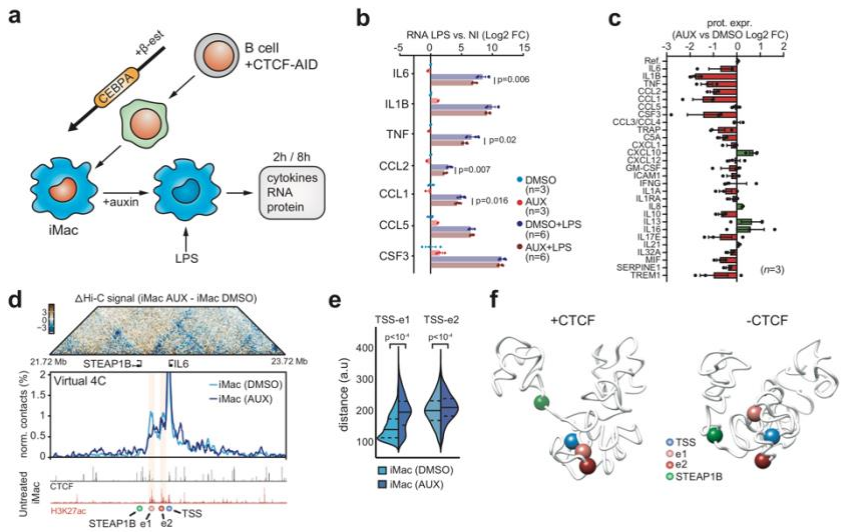


Fig. 4 | CTCF depletion attenuates the acute inflammatory response of iMacs to endotoxin. a, Schematic overview of the experiment. iMacs generated in the presence of CTCF were treated with either DMSO or auxin for 24 h followed by 2-8 h of LPS treatment and assayed for cytokine expression. **b**, qRT-PCR of selected

cytokine genes 2 h after LPS stimulation of iMacs pre-treated with DMSO or auxin. Error bars represent standard error; sample sizes (n) are indicated and represent biologically independent samples and *P* values derive from unpaired two-tailed *t* test. **c**, Secreted cytokine levels by iMacs treated with DMSO or auxin and stimulated for 8 h with LPS. Error bars represent standard error (n = 3 biologically independent samples) **d,** Top: Differential in situ Hi-C contact maps (10-kb resolution) at the *IL6* locus (chr7: 21.72-23.72 Mb) in iMacs generated in the presence of DMSO or auxin. Color scale represents differential contacts per 100,000 reads. The location of the *IL6* and *STEAP1B* genes is indicated; Middle part: Virtual 4C extracted from Hi-C data at the *IL6* locus in iMacs treated with DMSO or auxin, using the *IL6* transcription start site (TSS) as a viewpoint; Bottom part: browser snapshot showing CTCF and H3K27ac ChIP-seq signals. *IL6* enhancers (e1 and e2) are highlighted and the green, red and blue spheres represent the *STEAP1B* promoter, the *IL6* enhancers and the *IL6* TSS, respectively. **e,** Distance distribution between TSS and enhancer regions (n = 1,000 3D models based on Hi-C data of iMacs treated with DMSO or auxin). Median (solid line), first and third quartile (dashed line) are indicated (*P*, two-sided Komogorov-Smirnov test). **f**, 3D chromatin conformation model of the *IL6* locus in DMSO or auxin treated cells.

Our study has shown that the architectural protein CTCF is dispensable for the transdifferentiation of B cells into macrophages, while it is required for a full-blown inflammatory response. These findings indicate that CTCF-mediated genome topology, including TADs formed by cohesin-mediated loop extrusion, are not essential for developmental gene regulation but instead provide robustness and precision to an acute transcriptional response to bacterial endotoxins. Nevertheless, we cannot exclude that in other biological contexts, gene regulatory circuits especially dependent on CTCF-mediated genome topology might be more critically relevant. Importantly, our study uncouples the critical role of CTCF in cell proliferation from its role as genome organizer and transcriptional regulator and provides nuanced insights into the role of 3D chromatin organization for gene regulation. The observation that genome-wide CTCF is dispensable for a mammalian cell state transition significantly extends recent findings showing that CTCF or cohesin depletion during steady-state conditions, TAD rearrangements in flies or deletion of CTCF-mediated TAD boundaries in mice only caused minor changes in gene expression[13,14,39–41]. In addition, our findings indicate a critical role of 3D chromatin organization in providing an optimal response to external signals, in agreement with studies of developmentally regulated loci and nuclear hormone receptor signaling[39,42]. Future studies are required to assess whether this can be further generalized

to signaling responses during differentiation or development. In summary, we propose that cell fate transitions can occur in the absence of CTCF, while the effects of CTCF on genome topology are highly relevant for an acute transcriptional response. The observation that CTCF and global TAD organization are not strictly required for cell fate changes raises the possibility that lineage instructive transcription factors themselves shape multi-level topological genome dynamics relevant for major transcriptional rewiring.

Online content

Any methods, additional references, Nature Research report- ing summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0643-0.

References

1. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499–506 (2013).
2. Gorkin, D. U., Leung, D. & Ren, B. The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell* **14**, 762–775 (2014).
3. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164**, 1110–1121 (2016).
4. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
5. Furlong, E. E. M. & Levine, M. Developmental enhancers and chromosome topology. *Science (80-.).* **361**, 1341–1345 (2018).
6. Stadhouders, R., Filion, G. J. & Graf, T. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* **569**, 345–354 (2019).

7. Kim, S. & Shendure, J. Mechanisms of Interplay between Transcription Factors and the 3D Genome. *Mol. Cell* **76**, 306–319 (2019).
8. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
9. Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell* **48**, 471–484 (2012).
10. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
11. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).
12. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
13. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320 e24 (2017).
14. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944 e22 (2017).
15. Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56 (2017).
16. Haarhuis, J. H. I. *et al.* The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* **169**, 693-707.e14 (2017).
17. Heath, H. *et al.* CTCF regulates cell cycle progression of αβ T cells in the thymus. *EMBO J.* **27**, 2839–2850 (2008).
18. Arzate-Mejía, R. G., Recillas-Targa, F. & Corces, V. G. Developing in 3D: the role of CTCF in cell differentiation. *Development (Cambridge, England)* **145**, (2018).
19. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* **326**, 289–293 (2009).
20. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* (2015). doi:10.1016/j.cell.2015.04.004

21. Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900–10 (2015).
22. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.* (2015). doi:10.1073/pnas.1518552112
23. Narendra, V. *et al.* CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science (80-.).* (2015). doi:10.1126/science.1262088
24. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499–506 (2013).
25. Beagan, J. A. & Phillips-Cremins, J. E. On the existence and functionality of topologically associating domains. *Nat. Genet.* **52**, 8–16 (2020).
26. Rapino, F. *et al.* C/EBPalpha induces highly efficient macrophage transdifferentiation of B lymphoma and leukemia cell lines and impairs their tumorigenicity. *Cell Rep* **3**, 1153–1163 (2013).
27. Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
28. Stadhouders, R. *et al.* Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nat Genet* **50**, 238–249 (2018).
29. Natsume, T., Kiyomitsu, T., Saga, Y. & Kanemaki, M. T. Rapid Protein Depletion in Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors. *Cell Rep* **15**, 210–218 (2016).
30. Ouboussad, L., Kreuz, S. & Lefevre, P. F. CTCF depletion alters chromatin structure and transcription of myeloid-specific factors. *J. Mol. Cell Biol.* **5**, 308–22 (2013).
31. Nikolic, T. *et al.* The DNA-binding factor Ctcf critically controls gene expression in macrophages. *Cell. Mol. Immunol.* **11**, 58–70 (2014).
32. Cuartero, S. *et al.* Control of inducible gene expression links cohesin to hematopoietic progenitor self-renewal and differentiation. *Nat Immunol* **19**, 932–941 (2018).

33. Parelho, V. *et al.* Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422–433 (2008).
34. Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796–801 (2008).
35. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**, 919–922 (2016).
36. Faridi, M. H. *et al.* CD11b activation suppresses TLR-dependent inflammation and autoimmunity in systemic lupus erythematosus. *J. Clin. Invest.* **127**, 1271–1283 (2017).
37. Bell, A. C., West, A. G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**, 387–396 (1999).
38. Serra, F. *et al.* Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput. Biol.* **13**, (2017).
39. Despang, A. *et al.* Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).
40. Ghavi-Helm, Y. *et al.* Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.* (2019). doi:10.1038/s41588-019-0462-3
41. Williamson, I. *et al.* Developmentally regulated Shh expression is robust to TAD perturbations. *Dev.* **146**, (2019).
42. Le Dily, F. L. *et al.* Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev.* **28**, 2151–2162 (2014).

METHODS

Cell Culture

BLaER cell line[26] is derived from the RCH-ACV lymphoblastic leukemia cell line in which CEBPA fused with the estrogen receptor (ER) hormone binding domain and the GFP marker are expressed. BLaER cells and subclones were cultured in RPMI medium (Gibco, 22400089) supplemented with 10% fetal bovine serum (GIBCO, 10100147), 1% glutamine (GIBCO, 25030081), 1%

penicillin/streptomycin antibiotic (Thermo Fisher Scientific, 15140122), 550 µM β-mercaptoethanol (GIBCO, 31350010). Cells were maintained at a density of 0.1-6 × 10^6 cells/ml. Cells were checked for mycoplasma infection every month and tested negative. To induce transdifferentiation, BLaER cells were seeded at 0.3 million cells per ml in a culture medium supplemented with 100 nM β-estradiol, IL3 and CSF1 (100 ng/ml). iMacs were collected after 7 days of incubation. For auxin-inducible degradation, indole-3-acetic acid (IAA, a chemical analog of auxin) was added to the medium at 500 µM from a 1,000× stock diluted in dimethyl sulfoxide. Stocks were kept at 4ºC up to 4 weeks or -20ºC for long-term storage. For endotoxin stimulation, cells were treated with LPS (1 µg/ml) for 2 h to collect RNA or 8 h to collect supernatant.

Plasmid construction

The CTCF-mAID-mCherry targeting vectors were cloned by serial modification of the base vectors pMK292 (Addgene #72830) and pMK293 (Addgene #72831). Homology arms (HA) of the last exon of *CTCF* were synthesized (IDT®). The TIR1-AAVS1 donor vector pMK232 (Addgene #105924) and the pX330 vector expressing the sgRNA to target the *AAVS1* locus in human cells (Addgene #72833) were kindly provided by Masato Kanemaki[29]. *CTCF*-targeting sgRNAs were cloned in pX330 by annealing oligonucleotides caccgTGATCCTCAGCATGATGGAC and aaacGTCCATCATGCTGAGGATCAc.

Gene Targeting

For transfection, plasmids were prepared using Plasmid Midi Kit (Qiagen) followed by ethanol precipitation. Constructs were not linearized. The BLaER cell line was used to generate the parental line expressing the OsTIR1 enzyme. Transfection was carried out by electroporation (Amaxa Nucleofector, Lonza®), using Kit C and program X-001, according to manufacturer's instructions. One microgram of each plasmid (pMK232 and pX330-AAVS-sgRNA) was added per 100 µl of solution mix and 1 million cells. Eight millions of cells were transfected using the same conditions, and the day after the transfection, dead cells were eliminated by centrifugation and alive cells were pooled together. Three days after transfection, puromycin (1 µg/ml, Gibco A1113803) was added to the medium in order to select edited cells. Selection medium was changed each 2-3 days and the selection was performed for 10 days.

Single-cell sorting of resistant cells was performed and AAVS PCR genotyping allowed the selection of homozygous insertion of the TIR1 expression cassette at the *AAVS* locus. Several clones were selected and tested for TIR1 expression by qPCR allowing the selection of the clone with the most robust expression (cell line #2B10). This clone was used for the targeting of *CTCF*. Two runs of gene targeting were performed (the first using a Neomycin targeting plasmid and the second with a hygromycin targeting plasmid) to obtain homozygous recombined alleles. One µg of each plasmid (px330-mCherry-sgRNA; pHA-mAID-mCherry-Neo R or pHA-mAID-mCherry-Hygro R) was added per 100 µl of solution mix and 1 million cells. Eight millions of cells were transfected using the same conditions, and the day after the transfection, dead cells were eliminated by centrifugation and alive cells were pooled together. Three days after transfection, antibiotic was added to the medium in order to select edited cells (500 µg/ml G418, Life Technologies #11811031 and/or 100 µg/ml of Hygromycin B, GIBCO #10687010). Selection medium was changed each 2-3 days and the selection was performed for 17-20 days. Single-cell sorting of resistant cells expressing mCherry was performed and a genotyping PCR allowed the selection of homozygous mAID-CTCF targeted cells.

Flow Cytometry

BLaER cells and derived clones were resuspended in culture medium, spun down, and resuspended in 4% FBS-PBS and live (DAPI-negative) were sorted by live flow cytometry on a BD Influx™ instrument (BD Bioscience). For monitoring transdifferentiation, cells were subjected to a specific cell surface marker staining. Briefly, blocking was carried out for 10 min at room temperature (RT) using Human FcR Binding Inhibitor (1:20 dilution, eBiosciences, 16-9161-73) and cells were then stained with antibodies against CD19 (APC-Cy7 Mouse Anti-Human CD19, BD Pharmingen, 557791) and Mac-1 (APC Mouse Anti-Human CD11b/Mac-1, BD Pharmingen, 550019) at 4 ºC for 20 min in the dark. After washing, DAPI staining was performed just before analysis. For monitoring phagocytosis, cells were seeded at a density of 0.5 million/ml in medium and fluoresbrite carboxy bright blue beads (1 µm, Polysciences 17458) were added (300 beads/cell) and incubated 24 h before FACS analysis. Dissociation, wash, and flow buffers were supplemented with auxin, when appropriate, to

avoid re-expression of the CTCF-mAID-mCherry fusion. All the analyses were performed using the LSR Fortessa instrument (BD Biosciences). Data analysis was performed using FlowJo software.

Western Blots and antibody arrays

One million cells were centrifuged, washed with PBS 1× and lysis was performed in 30 µl of Laemmli buffer 1× (50 nM Tris-HCl pH6.8, 2% glycerol, 2% SDS, 0.01% 2-Mercaptoethanol, 0.05% Bromophenol blue). After heating at 95ºC for 10 min, the protein extracts (corresponding to 5×10^5 cells) were separated by electrophoresis in a 7.5% polyacrylamide gel (Bio-Rad #4561023) before transfer to nitrocellulose membrane. Membranes were blocked with 5% non-fat milk TBS-Tween medium (50 mM Tris, 150 mM NaCl, 0.1% Tween 20) for 1 h at RT. Incubation with primary antibody was performed at 4ºC shaking overnight (anti-CTCF, 07-729 Millipore; anti-α-tubulin, ab7291 Abcam; anti-mAID-tag MBL Life Science M214-3; 1:1,000 in 5% Milk in TBS-Tween). Membranes were washed with TBS-Tween (3 × 10 min) before secondary incubation with antibodies fused to HRP (goat anti-mouse IgG, Sigma Aldrich #A3682, dilution 1:5,000) for 1 h at RT. After 3 final washes, membranes were incubated in ECL™ Start Western Blotting Detection Reagent mix (Sigma Aldrich, GERPN3243) for 2 min at RT before development on X-ray film. Cytokine arrays (R&D ARY006) were performed following manufacturer's instructions using supernatant from iMacs collected 8 h after LPS stimulation (1 µg/ml). Antibody arrays were imaged using an Odyssey CLx instrument (LI-COR).

Immunofluorescence

iMacs were grown on glass-coverslips, fixed with 3% formaldehyde in PBS 1× for 10 min at RT. After washing with PBS, imaging was performed using a Leica TCS SPE inverted microscope. Images were post-processed using Fiji Is Just ImageJ (FIJI).

ChIP-seq

Cells were cross-linked for 10 min using 1% formaldehyde and quenched using a final concentration of 0.125 M glycin. Cell pellets were lysed by incubating 10 min on ice with 5 mM Pipes pH 8, 85 mM KCl, 0.5% IGEPAL, 1× protease inhibitor (Roche®). After centrifugation, pellets were incubated in 1% SDS, 10 mM EDTA pH 8, 50 mM Tris-HCl pH 8.1 and 1× PIC for 10 min on ice.

Chromatin was sheared on a Bioruptor pico sonicator (Diagenode) at 4ºC for 14 cycles of 30 sec ON and 30 sec OFF. After sonication, the solution was left on ice for 1 h to allow SDS precipitation and clarified by centrifugation at 16,000*g* at for 10 min at 4ºC. Supernatant was transferred in a new tube, 10% was saved as input and the rest was diluted to 1.2 ml with 1× cold IP buffer (Diagenode). 10 µg of anti-CTCF (Milipore, 07-729) was added followed by overnight incubation at 4ºC on a rotator. 42 µl of beads (Unblocked Protein A beads, kch-503-008, Diagenode) were used per IP after blocking them using 1% bovine serum albumin cold IP buffer for 15 min at 4ºC under rotation. Blocked beads were added to the chromatin solution and incubated 3 h at 4ºC with rotation. Beads were then collected by centrifugation for 2 min at 3,000 rpm at 4ºC and washed 3 times with cold IP buffer and 2 times with cold TE buffer (10 mM Tris pH 8, 1 mM EDTA). Beads were then eluted with freshly prepared elution buffer (1% SDS, 0.1 M $NaHCO_3$) and incubated 25 min at RT. The supernatant was transferred into a new tube and cross-linking was reversed by adding NaCl (final concentration 200 mM) and incubating overnight at 65ºC. Protein digestion was achieved by adding Tris pH 6.5 (40 mM), EDTA pH 8 (10 mM) and proteinase K (4 µg/µl) and incubating 1 h at 45ºC. DNA was then purified by phenol:chloroform:isoamyl alcohol (25:24:1) extraction. The entire DNA sample was used to construct Illumina sequencing libraries. Library preparation was performed using the NEBNext DNA Library Prep Kit (New England BioLabs) with 2 µl NEBNext adaptor in the ligation step. Libraries were amplified for 14 cycles with Herculase II Fusion DNA Polymerase (Agilent) and were purified/size-selected with Agencourt AMPure XP beads (> 200 bp). Libraries were sequenced on Illumina HiSeq2000 or NextSeq 500 instrument using 50 or 75 nucleotides paired-end mode, respectively.

Quantitative RT-PCR and RNA-seq

RNA was extracted with the miRNeasy mini kit (Qiagen) and quantified with a NanoDrop spectrophotometer. cDNA was produced with a High Capacity RNA-to-cDNA kit (Applied Biosystems) and was used for qRT–PCR analysis in triplicate reactions with SYBR Green QPCR Master Mix (Applied Biosystems). Oligonucleotide sequence are indicated in Supplementary Table 1. Libraries were prepared with an Illumina

TrueSeq Stranded total RNA Library Preparation Kit after Ribo-zero® depletion, and single-end sequencing (75 nt) was performed on an Illumina HiSeq2500 instrument.

ATAC-seq

ATAC-seq was performed as previously described[43]. Briefly, 5 million cells were harvested and treated with Nextera Tn5 Transposase (Illumina, FC-121-1030) for 45 min at 37°C. Library fragments were amplified using 1× NEBNext High-Fidelity 2× PCR Master Mix (NEB, M0541S) and 1.25 µM of custom Nextera PCR primers. PCR amplification was done with 11 cycles, determined by KAPA Real-Time Library Amplication Kit (Peqlab, KK2701) to stop prior to saturation. Then, the samples were purified using MinElute PCR Purification Kit (Qiagen, 28004) and with Agencourt AMPure XP beads (Beckman Coulter, A63881) in 3:1 ratio. The libraries were sequenced paired-end (50 bp) on a HiSeq2000 instrument.

***In situ* Hi-C library preparation**

In situ Hi-C was performed as previously described[12] with the following modifications: (i) two million cells were used as starting material; (ii) chromatin was initially digested with 100 U MboI (New England BioLabs) for 2 h, and then another 100 U (2 h incubation) and a final 100 U were added before overnight incubation; (iii) before fill-in with bio-dATP, nuclei were pelleted and resuspended in fresh 1× NEB2 buffer; (iv) ligation was performed overnight at 24°C with 10,000 cohesive end units per reaction; (v) de-cross-linked and purified DNA was sonicated to an average size of 300-400 bp with a Bioruptor Pico (Diagenode; 7 cycles of 20 sec on and 60 sec off); (vi) DNA fragment-size selection was performed only after final library amplification; (vii) library preparation was performed with an NEBNext DNA Library Prep Kit (New England BioLabs) with 3 µl NEBNext adaptor in the ligation step; (viii) libraries were amplified for 8–12 cycles with Herculase II Fusion DNA Polymerase (Agilent) and were purified/size-selected with Agencourt AMPure XP beads (> 200 bp). Hi-C library quality was assessed by low-coverage sequencing on an Illumina NextSeq500 instrument, after which every biological replicate (n = 2) was sequenced at high coverage on an Illumina HiSeq2500 instrument to obtain ~0.5 billion reads in total per time point per biological replicate.

In-situ Hi-C data processing and normalization

Hi-C data were processed using an in-house pipeline based on TADbit[38]. First, quality of the reads was checked using FastQC (http://www.bioinformatics.babraham.ac.uk /projects/fastqc/) to discard problematic samples and detect systematic artefacts. Trimmomatic[44] with the recommended parameters for paired end reads was used to remove adapter sequences and poor quality reads (ILLUMINACLIP:TruSeq3-PE.fa:2:30:12:1:true; LEADING:3; TRAILING:3; MAXINFO:targetLength:0.999; and MINLEN:36). For mapping, a fragment-based strategy as implemented in TADbit was used, which is similar to previously published protocols[45]. Briefly, each side of the sequenced read was mapped in full length to the reference genome (hg38, Dec 2017 GRCh38). After this step, if a read was not uniquely mapped, we assumed the read was chimeric due to ligation of several DNA fragments. We next searched for ligation sites and discarded reads in which no ligation site was found. Remaining reads were split as often as ligation sites were found. Individual split read fragments were then mapped independently. These steps were repeated for each read in the input FASTQ files. Multiple fragments from a single uniquely mapped read will result in as many contacts as possible pairs can be made between the fragments. For example, if a single read was mapped through three fragments, a total of three contacts (all-versus-all) was represented in the final contact matrix. We used the TADbit filtering module to remove non-informative contacts and to create contact matrices. The different categories of filtered reads applied are:

• self-circle: reads coming from a single restriction enzyme (RE) fragment and point to the outside.

• dangling-end: reads coming from a single RE fragment and point to the inside.

• error: reads coming from a single RE fragment and point in the same direction.

• extra dangling-end: reads coming from different RE fragments but are close enough and point to the inside. The distance threshold used was left to 500 bp (default), which is between percentile 95 and 99 of average fragment lengths.

• duplicated: the combination of the start positions and directions of the reads was repeated, pointing at a PCR artefact. This filter only removed extra copies of the original pair.

• random breaks: start position of one of the reads was too far from RE cutting site, possibly due to non-canonical enzymatic activity or random physical breaks. Threshold was set to 750 bp (default), > percentile 99.9. From the resulting contact matrices, low quality bins (those presenting low contacts numbers) were removed as implemented in TADbit's "filter_columns" routine. The matrices obtained were normalized for sequencing depth and genomic biases using OneD[46]. Then they were further normalized for local coverage within the region (expressed as normalized counts per thousand within the region) without any correction for the diagonal decay. For differential analysis, the resulting normalized matrices were directly subtracted from each other.

Identification of subnuclear compartments and topologically associating domains (TADs)

To segment the genome into A/B compartments, normalized Hi-C matrices at 100-kb resolution were corrected for decay as previously published, grouping diagonals when signal-to-noise was below 0.05[12]. Corrected matrices were then split into chromosomal matrices and transformed into correlation matrices using the Pearson product-moment correlation. The first component of a PCA (PC1) on each of these matrices was used as a quantitative measure of compartmentalization and AT content was used to assign negative and positive PC1 categories to the correct compartments. If necessary, the sign of the PC1 (which is randomly assigned) was inverted so that positive PC1 values corresponded to A compartment regions and vice versa for the B compartment. Significant differences of PC1 values between conditions were calculated using two-sided Wilcoxon rank-sum tests. Normalized contact matrices at 50-kb resolution were used to define TADs, using a previously described method[27] with default parameters. First, for each bin, an insulation score was obtained based on the number of contacts between bins on each side of a given bin. Differences in insulation score between both sides of the bin were computed and borders were called searching for minima within the insulation score[27]. This procedure resulted in a set of borders for each time point and replicate. Between replicates, overlapping or borders distant of less than 1 bin were merged to obtain a list of conserved borders for each time point. Conserved borders overlapping or distant of less than 1 bin among each time point were considered as stable while the others were considered as

dynamic. Significant differences of insulation scores between conditions were calculated using two-sided Wilcoxon rank-sum tests.

Inter- and intra-compartment strength measurements

We followed a previously reported strategy to measure overall interaction strengths within and between A and B compartments[15]. Briefly, we based our analysis on the 100-kb bins showing the most extreme PC1 values, discretizing them by percentiles and taking the bottom 20% as B compartment and the top 20% as A compartment. We classified each bin in the genome according to PC1 percentiles and gathered contacts between each category, computing the $\log_2$ enrichment over the expected counts by distance decay. Finally, we summarize each type of interaction (A-A, B-B and A-B/B-A) by taking the median values of the $\log_2$ contact enrichment.

Meta-analysis of borders

To study the behavior of TAD borders all TADs of sizes ranging from 0.5 to 1.5 Mb were selected. Then we defined a flanking region of 1 Mb around the border and gathered the observed and expected (by distance decay) matrix counts. Setting up their relative position to the corresponding border, the matrices were stacked to obtain a meta-contact matrix around TAD borders for each condition. This information was summarized by comparing the average $\log_2$ fold change of contact enrichment between bins inside and outside TAD.

Enhancer-promoter intra-TAD contacts analysis

By using Hi-C matrices at 5-kb resolution, we focused on TADs containing enhancers and promoters. Each bin was labelled as part of an enhancer, promoter or "others" if they did not belong to previous types. Then the observed contacts were gathered between the different types of bins within their TAD and expected contact frequencies were computed based on the genomic distance that separate each pair (the expected distance decay was calculated excluding entries outside TADs). Then a liner mixed models including TAD ID as random effect was used to estimate the quantities of interest. Results are expressed as $\log_2$ of the ratio observed on expected frequencies of contacts.

Long-range interactions between enhancers and promoters
Hi-C matrices were generated at 10-kb resolution using HiCExplorer[47] and long range interactions (5-10 Mb) between promoters and enhancers activated or inactivated were computed using the HiCExplorer tool hicAggregateContacts.

Meta-analysis of cohesin loops
Hi-C matrices in cool format were used to generate genome-wide aggregate plots at SMC1-bound loops detected by HiChIP[35]. We used *coolpup.py*[48] to pile-up normalized Hi-C signals at a 10-kb resolution at SMC1-bound loops previously identified[35], and plotted 100 kb upstream and downstream of the SMC1 anchor coordinates.

Virtual 4C analysis
Hi-C matrices for virtual 4C profiles were further smoothed using a focal (moving window) average of one bin. The profiles were generated from these normalized matrices and correspond to histogram representation of the lines of the matrices containing the baits (therefore expressed as counts per thousand of normalized reads within the region depicted).

Gene expression analysis using RNA-seq data
Reads were mapped using STAR[49] (standard options) and the Ensembl human genome annotation (GRCh38v27). Gene expression was quantified using STAR (--quantMode GeneCounts). Batch effects were removed using ComBat function from sva R package (v3.22). Sample scaling and statistical analysis were performed using the R package DESeq2[50] (R 3.3.2 and Bioconductor 3.0). Genes changing significantly at any time point were identified using the nbinomLRT test (FDR < 0.01) and fold-change > 2 between at least two time points. Log_2-vsd (variance stabilized DESeq2) counts were used for further analysis unless stated otherwise. To compare expression of various set of genes the data were mean-centered log-transformed and significant differences were calculated using two-sided Wilcoxon rank-sum tests.

Chromatin accessibility analysis using ATAC-seq data
Reads were mapped to the UCSC human genome build (hg38) using Bowtie2[51] with standard settings. Reads mapping to multiple locations in the genome were removed using SAMtools[52]; PCR

duplicates were filtered using Picard (http://broadinstitute.github.io/picard). Bam files were parsed to deepTools[53] for downstream analyses and browser visualization. Peaks in ATAC-seq signal were identified using MACS2[54] (callpeak --nolambda --nomodel -g hs -q 0.01).

ChIP-seq data analysis

Reads were mapped and filtered as described for ATAC-seq. CTCF peaks were identified using MACS2[54] with the *narrowpeaks* option. CTCF peaks not called in both independent biological replicates were excluded in all subsequent analyses. Coverage of CTCF peaks per TAD border was computed using BEDTools[55]. H3K27ac coverage and CTCF binding heatmaps were performed using deepTools[53].

DNA motif analysis

ATAC-seq peaks specific to the TAD borders were identified using bedtools[55]. DNA motif analysis of the ATAC-seq peaks were analyzed using HOMER[56] (*findMotifs.pl*) and the Homer motif results were shown. It uses ZOOPS scoring (zero or one occurrence per sequence) coupled with hypergeometric test to determine motif enrichment and statistical significance.

Identification of dynamic regulatory regions

Intersecting ATAC-seq peaks with H3K4me1 peaks allowed the identification of 63,665 enhancers, while the overlap with transcription start sites (TSS) revealed 24,932 promoters (**Extended Data Fig. 2g**). The intensity of H3K27ac signals at these regions was quantified using the Diffbind R package (v2.2.12) to define activated and inactivated regions from 0 h to 168 h. Differences were computed with using DBA_DESEQ2 method and -filter for significance was set at fold change > 2 and FDR < 0.05) as previously described[57]. This analysis allowed to profile 29,711 dynamic enhancers and 8,439 dynamic promoters of which about half became activated and the other half inactivated (**Extended Data Fig. 2h**), also reflected by the expression of the associated genes (**Extended Data Fig. 2i**).

3D modeling and analysis

The processed Hi-C datasets were binned at 10-kb resolution and then normalized using OneD[46]. Then, we defined the regions to be modelled given the genomic context around the enhancer and promoters of interest by following the steps: (i) select key elements contained in the region (i.e., enhancers and promoters); (ii) retrieve the top 5% interactors of each of these elements; (iii) build a network joining the key elements with their retrieved top 5% interactors, and the top 5% interactors among them in the cases where this interaction (interactor with interactor) was present in the top 5% of at least one of them. Added the edge twice if it was in the top 5% of both members; (iv) group the networks allowing a genomic distance gap of 50 kb and filtered out the groups in which the ratio between (number of edges) / (Number of nodes) was smaller than 5; and (v) for each of the regions, ensure that the modelled region contained most of the nodes (genomic coordinate from one bin start until end) appearing in the groups that passed the previous filter. Once regions were selected, normalized interaction matrices were modelled as previously described[58] using TADdyn[59], a molecular dynamic-based protocol implemented in the TADbit software[38]. Similarly to TADbit, TADdyn generates models using a restraint-based approach, in which experimental frequencies of interaction are transformed into a set of spatial restraints[60]. A total of 1,000 models were generated for each genomic region and cell type. Contact maps generated from the ensemble models highly correlated with the input Hi-C normalized interaction matrices. Each ensemble of models was next clustered based on structural similarity as implemented in TADbit[38]. The absence of major structural differences between clusters prompted us to use all them in further analysis. Next, TADbit was used to measure the following features of the models: (i) distance between particles containing genomic regions of interest in the model ensemble; distances distribution between selected pairs or particles, and (iii) significant differential distance distributions assessed by two-sample Kolmogorov-Smirnov statistic. Finally, model images were generated with Chimera[61].

Statistics and reproducibility

RNA-seq and in situ Hi-C data throughout the paper were generated by analysis of 2 biologically independent samples from 2 transdifferentiation experiments. Representative data are shown

only if results were similar for both biologically independent replicates. All box plots depict the first and third quartiles as the lower and upper bounds of the box, with a band inside the box showing the median value and whiskers representing 1.5× the interquartile range. Wilcoxon rank-sum tests were performed with the wilcox.test() function in R in a two-sided manner. Student's *t* tests were performed with the t.test() function in R in an unpaired and two-sided fashion with (n-2) degrees of freedom. Kolmogorov-Smirnov tests were performed in a two-side manner using the module scipy.stats.ks_2sam in the SciPy software.

Reporting Summary

Further information on research design is available in the Life Sciences Reporting Summary linked to this article.

Data availability

The Hi-C, RNA-seq, CTCF-ChIP-seq, ATAC-seq datasets generated and analyzed for the current study are available in the Gene Expression Omnibus (GEO) database under accession number GSE140528. ATAC-seq and CEBPA ChIP-seq datasets used in the current study are available in the GEO database under accession number GSE131620. The H3K27ac and H3K4me1 ChIP-seq datasets used in this study are available in ArrayExpress database under accession number E-MTAB-9010.

Methods-only-References

43. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
44. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
45. Ay, F. *et al.* Identifying multi-locus chromatin contacts in human cells using tethered multiple 3C. *BMC Genomics* **16**, (2015).
46. Vidal, E. *et al.* OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Res* **46**, e49 (2018).

47. Ramírez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, (2018).
48. Flyamer, I. M., Illingworth, R. S. & Bickmore, W. A. Coolpup.py: versatile pile-up analysis of Hi-C data. *Bioinformatics* (2020). doi:10.1093/bioinformatics/btaa073
49. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
50. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, (2014).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
52. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. DeepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, (2014).
54. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
55. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
56. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).
57. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
58. Miguel-Escalada, I. *et al.* Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nat. Genet.* (2019). doi:10.1038/s41588-019-0457-0
59. Stefano, M. Di *et al.* Dynamic simulations of transcriptional control during cell reprogramming reveal spatial chromatin caging. *bioRxiv* 1–29 (2019). doi:10.1101/642009
60. Baù, D. & Marti-Renom, M. A. Genome structure determination via 3C-based data integration by the

Integrative Modeling Platform. *Methods* **58**, 300–306 (2012).
61. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

Acknowledgments

We thank M. T. Kanemaki for the degron plasmids; Roderic Guigó's laboratory, and Silvia Pérez-Lluch in particular, for the H3K27ac and H3K4me1 ChIP-seq, produced in the framework of the RNA-MAPS project (ERC-2011-AdG-294653-RNA-MAPS); Y. Cuartero for help with sequencing and CTCF ChIP-seq; C. Segura for help with immunofluorescence microscopy; the CRG Genomics and flow cytometry core facilities and the CRG-CNAG Sequencing Unit for sequencing; and members of T.G.'s laboratory for discussions. This work was supported by the European Research Council under the 7th Framework Programme FP7/2007-2013 (ERC Synergy Grant 4D-Genome, grant agreement 609989 to T.G., M.A.M.-R.), the Ministerio de Educación y Ciencia (SAF.2012-37167 to T.G. and BFU2017-85926-P to M.A.M.-R.), the AGAUR to T.G., the Marató TV3 (201611) to M.A.M.-R.). P.C. was supported by the Deutsche Forschungsgemeinschaft (SFB860, SPP1935, EXC 2067/1- 390729940), the European Research Council (advanced investigator grant TRANSREGULON, grant agreement no. 693023), and the Volkswagen Foundation. G.S. was supported by a Marie Skłodowska-Curie fellowship (H2020-MSCA-IF-2016, miRStem) and by the 'Fundación Científica de la Asociación Española Contra el Cáncer'. T.V.T. was supported by Juan de la Cierva postdoctoral fellowship (MINECO; FJCI-2014-22946). B.B. was supported by the fellowship 2017FI_B00722 from the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement (Generalitat de Catalunya) and the European Social Fund (ESF). R.S. was supported by a Netherlands Organisation for Scientific Research Veni fellowship (91617114) and an Erasmus MC Fellowship. We also acknowledge support from 'Centro de Excelencia Severo Ochoa 2013-2017' (SEV-2012-0208), the Spanish Ministry of Science and Innovation to the EMBL partnership and the CERCA Program Generalitat de Catalunya to the CRG, as well as the support of the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III, the

Generalitat de Catalunya through Departament de Salut and Departament d'Empresa i Coneixement, and co-financing by the Spanish Ministry of Science and Innovation with funds from the European Regional Development Fund (ERDF) corresponding to the 2014-2020 Smart Growth Operating Program to CNAG.

Author contributions

G.S., R.S. and T.G. conceived the study and wrote the manuscript with input from all coauthors. G.S. performed molecular biology, RNA-seq, ChIP-seq, and in situ Hi-C experiments. T.V.T., J.C. and A.A. performed ChIP-seq and ATAC-seq. G.S., E.V., M.V.-C., J.M.-E. and B.B. performed bioinformatic analyses. G.S., E.V., M.V.-C., J.M.-E., and R.S. integrated and visualized data. G.S. and M.B. performed the CTCF-degron CRISPR targeting and S.C. performed cytokine arrays. G.S. performed the transdifferentiation experiments with help from M.B. and C.B.; F.L.D., P.C., M.A.M.-R. and R.S. provided valuable advice and T.G. supervised the research.

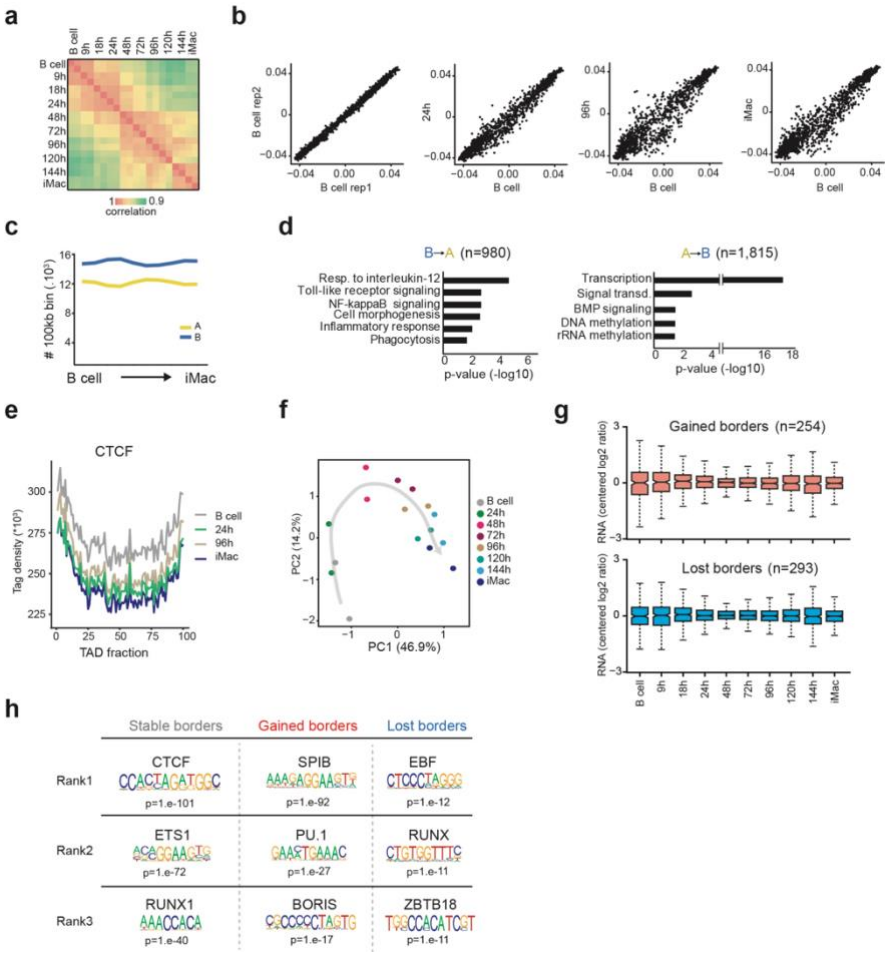
Competing interests Statement

The authors declare no competing interests

Additional information

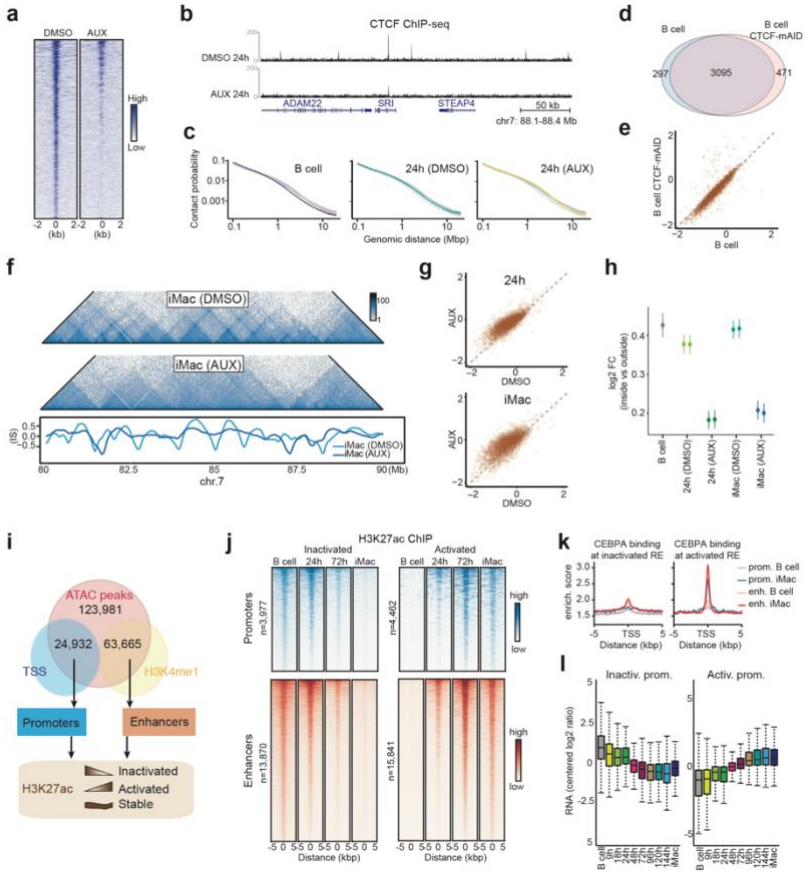
Extended data is available for this paper
Supplementary information is available for this paper
Correspondence and requests for materials should be addressed to G.S., R.S. or T.G.

Supplementary Figures



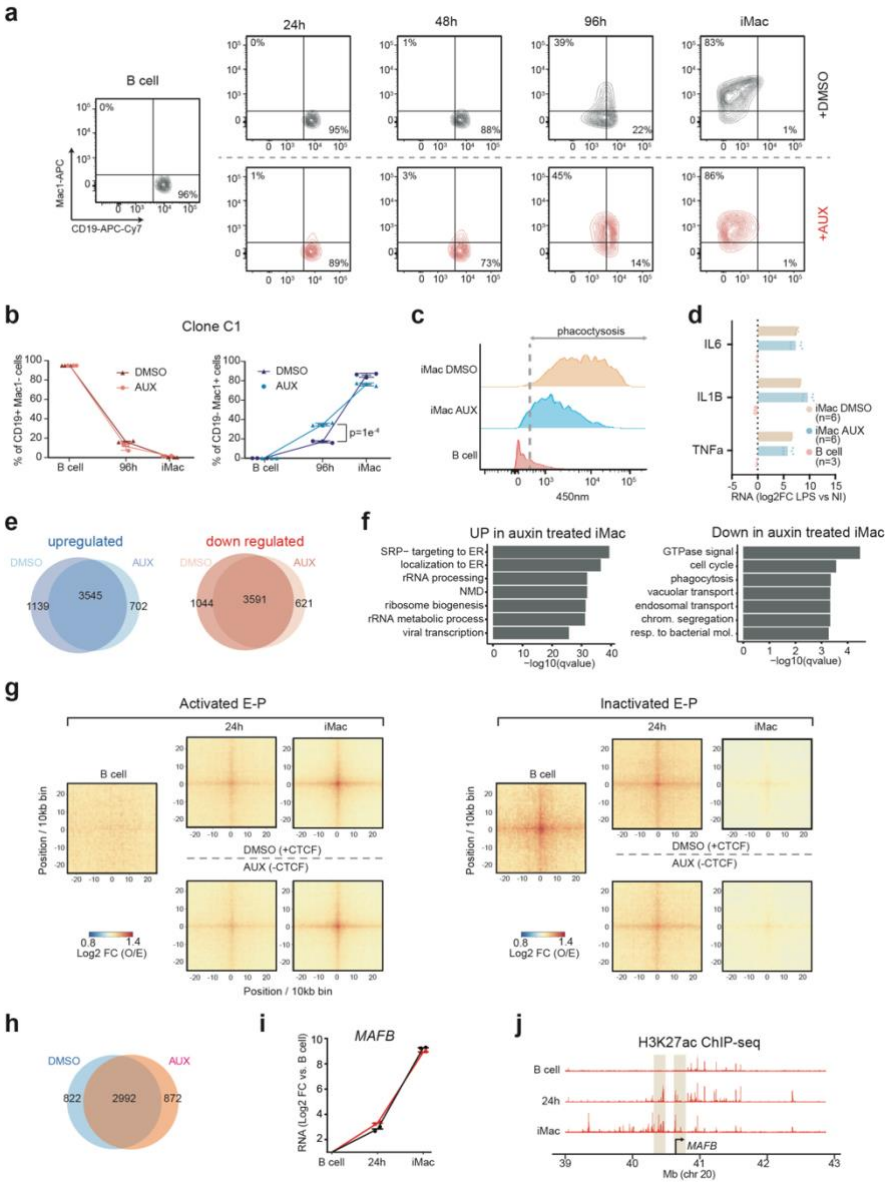
Extended Data Fig. 1 | Characterization of chromatin compartmentalization and TAD dynamics during transdifferentiation. **a**, Genome-wide Pearson correlation matrix between PC1 values of Hi-C samples at different time points. **b**, Scatter plots of PC1 values (n = 1,332 100-kb bins) showing changes relative to initial B cell genome compartmentalization for chromosome 12. **c**, Line chart depicting fractions of the genome assigned to A or B compartments at 10 time points during transdifferentiation. Y-axis represents the number of 100-kb bins. **d**, Gene ontology analysis of genes in regions switching from B to A (n = 980 genes) or A to B (n = 1,815 genes) compartments (*P* values, FDR corrected Fisher test). **e**, CTCF binding signal at TADs, normalized for TAD size in samples at various transdifferentiation time points. **f**, PCA of insulation score values at TAD borders during transdifferentiation (n = 4,006 TAD borders). Grey arrow depicts an averaged trajectory. **g**, RNA expression of genes at TAD borders gained (n = 254 genes) or lost (n = 293 genes) during transdifferentiation. All box plots depict the first and third quartiles as the lower and upper bounds of the box, with a thicker

band inside the box showing the median value and whiskers representing 1.5x the interquartile range. **h**, Homer DNA motif analysis at ATAC-seq peaks detected at stable (n = 2,044), gained (n = 591) or lost (n = 135) TAD borders (*P* values are calculated using hyper-geometric statistical tests).



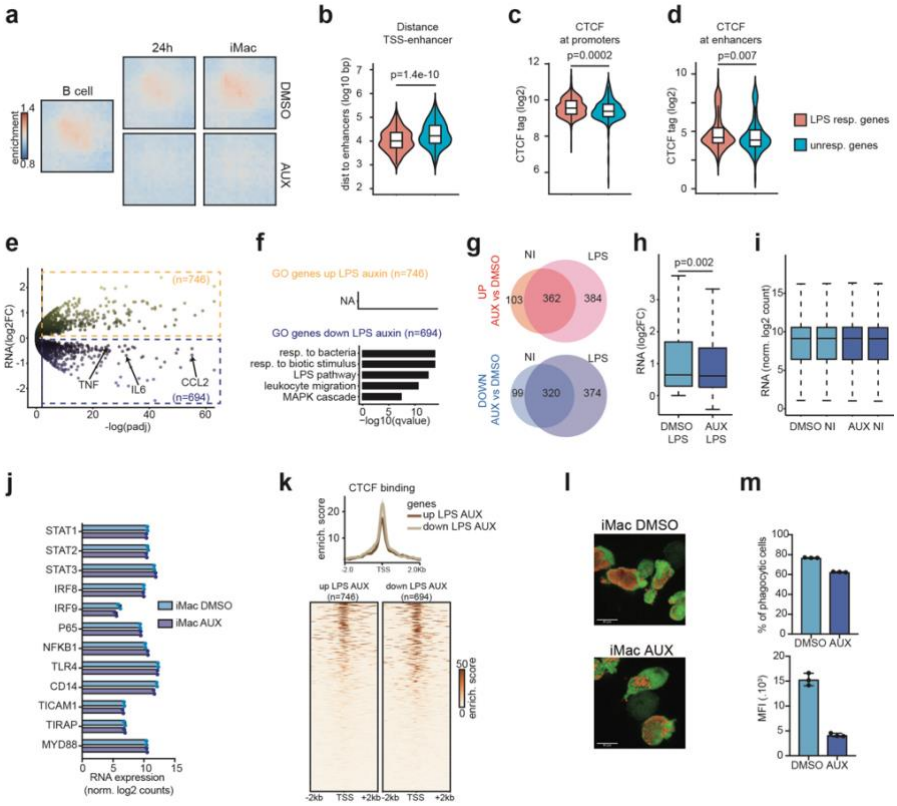
Extended Data Fig. 2 | Molecular characterization of CTCF-mAiD BLaER cells during transdifferentiation. a, Heatmap of CTCF binding signal at CTCF ChIP-seq peaks detected in untreated CTCF-mAID cells. **b**, Browser snapshot showing CTCF binding loss upon 24 h of auxin treatment. **c**, Overall scaling of Hi-C contact frequency as a function of genomic distance in cells treated with DMSO or auxin. **d**, Venn diagram showing the overlap of TAD borders detected in B cells and in CTCF-AID B cells. **e**, Scatter plots comparing insulation scores at TAD borders at B cell and at CTCF-AID B cells. Lower values indicate stronger insulation. **f**, Top: Representative in situ Hi-C contact maps (20-kb resolution) of iMacs obtained after treatment with DMSO or auxin. Color scale represents the normalized number of contacts. Bottom: plots of the corresponding insulation scores for each bin within the 10-Mb region shown. **g**, Scatter plots comparing insulation scores at TAD borders at 24 hpi and at the iMac stage after DMSO or auxin treatment. Lower values indicate stronger insulation. **h**, Contact enrichment of interactions inside TADs versus outside TADs at the indicated time points for B cell (n = 1), DMSO- (n = 2) or auxin-treated cell (n = 2) biologically independent samples. Dots represent point estimates and bars (wide and narrow) indicate confidence intervals (50% and 95 %, respectively) for the log2 fold changes. All estimations are computed using all 9 samples in a single linear mixed model. **i**, Outline of strategy

used to identify dynamic promoters and enhancers during transdifferentiation. Numbers of ATAC-seq peaks intersecting with TSS (promoters) and H3K4me1 peaks (enhancers) are indicated. **j**, H3K27ac decoration at dynamic promoters and enhancers that become either inactivated or activated. **k**, CEBPA binding at activated and inactivated regulatory elements (RE) in B cell and iMacs. **l**, RNA expression of genes associated with inactivated (n = 1,259) and activated (n = 1,421) promoters during transdifferentiation. All box plots depict the first and third quartiles as the lower and upper bounds of the box, with a thicker band inside the box showing the median value and whiskers representing 1.5x the interquartile range.



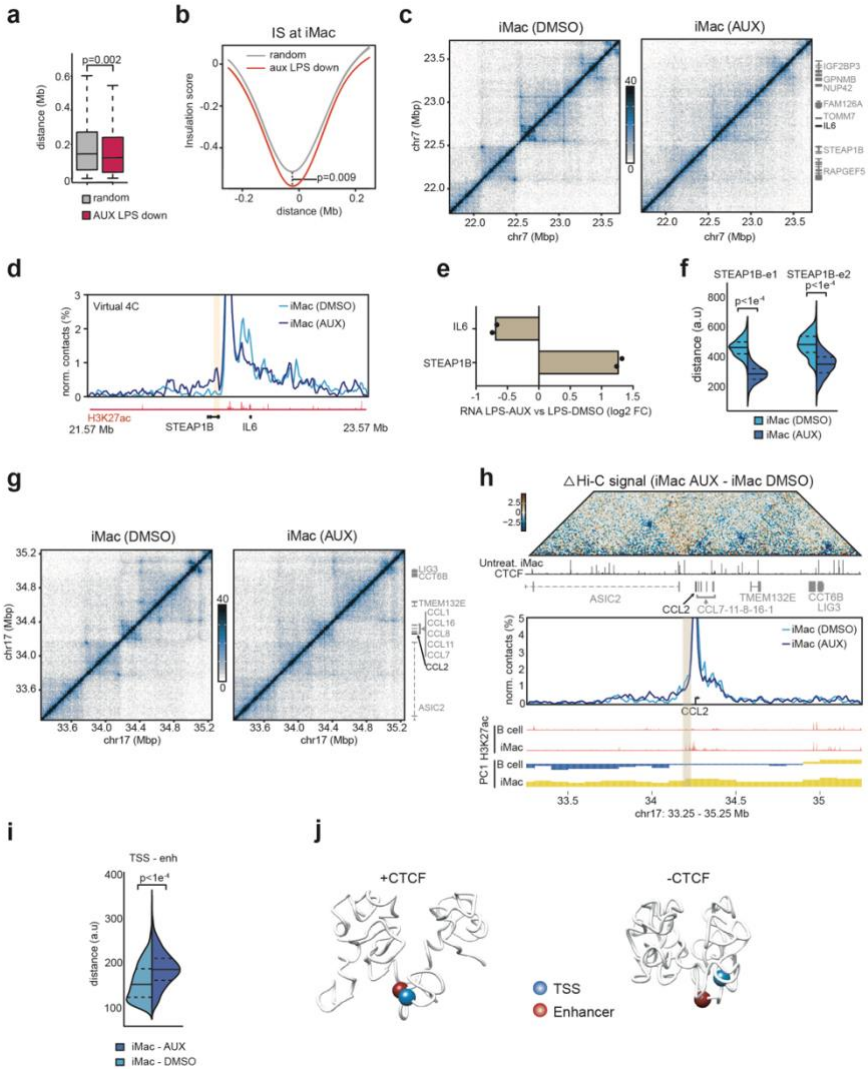
Extended Data Fig. 3 | CTCF depletion does not impair transdifferentiation or long-range enhancer-promoter contact dynamics. a, Representative flow cytometry analysis of CD19 and Mac-1 marker expression during transdifferentiation of CTCF-mAID B cells treated with DMSO or auxin. The experiment was repeated 3 times with similar results. **b**, Transdifferentiation kinetics of CTCF-mAID B cells (clone C1) in the presence of DMSO or auxin analysed at 0, 96 and 168 hpi by flow cytometry for CD19 and Mac-1 expression (n = 3 biologically independent samples). Centre indicates mean, error bars show standard deviation and *P* unpaired two-tailed t-test. **c**, Phagocytosis assay of iMacs analyzed by flow cytometry showing uptake of blue fluorescent beads. The experiment was

repeated 3 times with similar results. **d**, RNA expression measured by qRT-PCR of cytokines in noninduced (NI) or 2h LPS-induced iMacs DMSO (n = 6), iMacs AUX (n = 6) or B cells (n = 3). Mean values are shown, error bars represent standard error and *n* represents biologically independent samples. **e**, Venn diagram showing the overlap of genes upregulated (left) and downregulated (right) in iMacs after transdifferentiation in the presence of DMSO or auxin based on RNA-Seq (n = 2 biologically independent samples). **f**, Gene ontology analysis of genes specifically upregulated (n = 419) and downregulated (n = 744) specifically in CTCF-depleted iMacs (q-value, FDR corrected Fisher exact test). **g**, Aggregate metaplots (10-kb resolution) depicting long range (5–10-Mb) interaction frequencies between enhancers and promoter (E-P) during transdifferentiation. Area shown is centered on enhancers or promoters ± 250-kb). **h**, Venn diagram showing the number of switching compartment regions (100-kb bins) during transdifferentiation in presence of DMSO or auxin. **i**, Expression of *MAFB* during transdifferentiation with or without CTCF, as measured by RNA-seq (n = 2 biologically independent samples, lines connect mean values). **j**, Enhancer activity at the *MAFB* locus during transdifferentiation. Browser snapshot showing H3K27ac ChIP-seq profiles of a 4-Mb domain surrounding the *MAFB* locus. The enhancer and the promoter shown in Fig. 3g are highlighted in light brown.



Extended Data Fig. 4 | CTCF depletion in iMacs attenuates the acute inflammatory response to endotoxins. a, Genome-wide aggregation of normalized Hi-C signal anchored at cohesin loops during transdifferentiation with DMSO or auxin. **b**, Distance distribution between enhancers and TSS of genes responsive (n = 378) or unresponsive (n = 380) to LPS (*P*, two-sided Wilcoxon rank-sum test). **c**, CTCF enrichment at promoters and enhancers **d**, of genes responsive (n = 378) or unresponsive (n = 378) to LPS (*P*, two-sided Wilcoxon rank-sum test). **e**, Differential gene expression between LPS-induced iMac treated with auxin or DMSO (n = 2 biologically independent samples, *P*-adj two-tailed likelihood ratio test followed by FDR correction). **f**, Gene ontology analysis of the significantly ($p < 0.01$) upregulated (n = 746) and downregulated (n = 694) genes in LPS-induced iMacs treated with auxin compared to DMSO (q-value, FDR-corrected Fisher exact test). **g**, Overlap of upregulated (top) and downregulated (bottom) genes (AUX vs DMSO) between non-induced iMacs (NI) and iMacs treated with LPS (LPS). **h**, LPS-upregulated genes in iMacs exposed to DMSO compared to auxin (n = 2,470 genes, *P* two-sided Wilcoxon rank-sum test). **i**, RNA expression in non-induced (NI) iMacs of genes upregulated after LPS stimulation of DMSO treated iMacs (n = 2,470). **j**, RNA expression of key transcription factors and receptors of the LPS signalling pathway (n = 2 biologically independent samples). **k**, CTCF binding at promoters of genes deregulated in LPS-induced iMacs treated with auxin as compared to DMSO. **l**, Micrographs show uptake of fluorescent beads (shown in red) by iMacs treated with DMSO or auxin (Scale bar

represents 10 μm). The experiment was repeated 3 times with similar results. **m**, Quantification of phagocytosis assay. Upper panel shows percentage of cells with bead uptake; lower panel shows mean fluorescent intensity (MFI). Bars represent mean values of n = 3 biologically independent samples and error bars denote standard deviation. All box plots depict the first and third quartiles as the lower and upper bounds of the box, with a thicker band inside the box showing the median value and whiskers representing 1.5x the interquartile range.



Extended Data Fig. 5 | CTCF depletion in iMacs impairs 3D chromatin organization at inflammatory response gene loci. a, Distance distribution between promoters and their closest TAD borders of genes downregulated in auxin treated iMacs after LPS induction (as compared to iMacs exposed to DMSO) or for a random set of genes with a similar size (n = 687) (*P*, two-sided Wilcoxon rank-sum test). Box plots depict the first and third quartiles as the lower and upper bounds of the box, with a thicker band inside the box showing the median value and whiskers representing 1.5x the interquartile range. **b**, Average insulation scores of TAD borders closest to genes downregulated in auxin treated iMacs after LPS induction or closest to a random set of genes with a similar size (n = 687). Area shown is centered on boundary regions ± 250-kb (*P*, two-sided Wilcoxon rank-sum test). **c**, Hi-C maps (10-kb resolution) at the *IL6* locus. Color scale represents the normalized number of contacts and the genes within the locus are indicated on the right. **d**,

Virtual 4 C of iMacs treated with DMSO (dark blue) or auxin (light blue) using *IL6* enhancer 1 (e1) as viewpoint; Browser snapshot of H3K27ac ChIP-seq signal is shown and the *STEAP1B* promoter is highlighted. **e**, Differential expression of *IL6* and *STEAP1B* in LPS-induced iMacs treated with auxin as compared to DMSO (bars represent the mean values of n = 2 biologically independent samples). **f**, Distance distribution between STEAP1B promoter and IL6 enhancer regions (n = 1,000 models). Median (solid line), first and third quartile (dashed line) are indicated (*P*, two-sided Komogorov-Smirnov test). **g**, Hi-C maps (10-kb resolution) at the *CCL2* locus in iMacs generated in the presence of DMSO or auxin. Color scale represents the normalized number of contacts. Genes within the locus are indicated on the right. **h**, Top: Differential Hi-C maps of the *CCL2* locus (10-kb resolution) in iMacs generated in the presence of DMSO or auxin; CTCF ChIP-seq signal and gene positions are shown below the Hi-C map. Middle: Virtual 4 C of the *CCL2* locus of iMacs treated with DMSO (dark blue) or auxin (light blue), using *CCL2* promoter as viewpoint. Bottom: browser snapshots showing H3K27ac ChIP-seq and PC1 A/B compartment tracks. The *CCL2* enhancer is highlighted. **i**, Distance distribution between CCL2 TSS and enhancer regions (n = 1,000 models). Median (solid line), first and third quartiles (dashed line) are indicated (*P*, two-sided Komogorov-Smirnov test). **j**, 3D model of the *CCL2* locus in DMSO or auxin treated iMacs.

Supplementary Table 1. Sequences of oligonucleotides used for qRT-PCR

Oligonucleotide	Sequence
IL6_Forward	AGTGAGGAACAAGCCAGAGC
IL6_Reverse	ATTTGTGGTTGGGTCAGGGG
IL1B_Forward	CGCCAGTGAAATGATGGCTT
IL1B_Reverse	ATCCAGAGGGCAGAGGTCC
TNF_Forward	ACTTTGGAGTGATCGGCCC
TNF_Reverse	CATTGGCCAGGAGGGCATT
CCL2_Forward	ATCAATGCCCCAGTCACCTG
CCL2_Reverse	TCTCCTTGGCCACAATGGTC
CCL1_Forward	CGGAGCAAGAGATTCCCCTG
CCL1_Reverse	TGCCTCTGAACCCATCCAAC
CCL5_Forward	TCATTGCTACTGCCCTCTGC
CCL5_Reverse	CACACTTGGCGGTTCTTTCG
CSF3_Forward	GAGTGTGCCACCTACAAGCT
CSF3_Reverse	CCGCTATGGAGTTGGCTCAA
GAPDH_Forward	CAGCCTCAAGATCATCAGCA
GAPDH_Reverse	TGTGGTCATGAGTCCTTCCA

REFERENCES

Abbas, A., X. He, J. Niu, B. Zhou, G. Zhu, T. Ma, . . . J. Zeng (2019). "Integrating Hi-C and FISH data for modeling of the 3D organization of chromosomes." Nat Commun **10**(1): 2049.

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

Allahyar, A., C. Vermeulen, B. A. M. Bouwman, P. H. L. Krijger, M. Verstegen, G. Geeven, . . . W. de Laat (2018). "Enhancer hubs and loop collisions identified from single-allele topologies." Nat Genet **50**(8): 1151-1160.

Andersson, R. and A. Sandelin (2020). "Determinants of enhancer and promoter activities of regulatory elements." Nat Rev Genet **21**(2): 71-87.

Anil, A., R. Spalinskas, Ö. Åkerborg and P. Sahlén (2018). "HiCapTools: a software suite for probe design and proximity detection for targeted chromosome conformation capture applications." Bioinformatics (Oxford, England) **34**(4): 675-677.

Ardakany, A. R., F. Ay and S. Lonardi (2019). "Selfish: discovery of differential chromatin interactions via a self-similarity measure." Bioinformatics (Oxford, England) **35**(14): i145-i153.

Arrastia, M. V., J. W. Jachowicz, N. Ollikainen, M. S. Curtis, C. Lai, S. A. Quinodoz, . . . R. F. Ismagilov (2020). "A single-cell method to map higher-order 3D genome organization in thousands of individual cells reveals structural heterogeneity in mouse ES cells." bioRxiv: 2020.2008.2011.242081.

Bartke, T., M. Vermeulen, B. Xhemalce, S. C. Robson, M. Mann and T. Kouzarides (2010). "Nucleosome-interacting proteins regulated by DNA and histone methylation." Cell **143**(3): 470-484.

Baù, D., A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, . . . M. A. Marti-Renom (2011). "The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules." Nat Struct Mol Biol **18**(1): 107-114.

Bauman, J. G., J. Wiegant, P. Borst and P. van Duijn (1980). "A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA." Exp Cell Res **128**(2): 485-490.

Beagrie, R. A., A. Scialdone, M. Schueler, D. C. Kraemer, M. Chotalia, S. Q. Xie, . . . A. Pombo (2017). "Complex multi-enhancer contacts captured by genome architecture mapping." Nature **543**(7646): 519-524.

Beagrie, R. A., C. J. Thieme, C. Annunziatella, C. Baugher, Y. Zhang, M. Schueler, . . . A. Pombo (2020). "Multiplex-GAM: genome-wide identification of chromatin contacts yields insights not captured by Hi-C." bioRxiv: 2020.2007.2031.230284.

Beliveau, B. J., A. N. Boettiger, G. Nir, B. Bintu, P. Yin, X. Zhuang and C. T. Wu (2017). "In Situ Super-Resolution Imaging of Genomic DNA with OligoSTORM and OligoDNA-PAINT." Methods Mol Biol **1663**: 231-252.

Beliveau, B. J., E. F. Joyce, N. Apostolopoulos, F. Yilmaz, C. Y. Fonseka, R. B. McCole, . . . C. T. Wu (2012). "Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes." Proc Natl Acad Sci U S A **109**(52): 21301-21306.

Ben Zouari, Y., A. M. Molitor, N. Sikorska, V. Pancaldi and T. Sexton (2019). "ChiCMaxima: a robust and simple pipeline for detection and visualization of chromatin looping in Capture Hi-C." Genome Biol **20**(1): 102.

Bendandi, A., S. Dante, S. R. Zia, A. Diaspro and W. Rocchia (2020). "Chromatin Compaction Multiscale Modeling: A Complex Synergy Between Theory, Simulation, and Experiment." Front Mol Biosci **7**: 15.

Bernstein, B. E., E. Birney, I. Dunham, E. D. Green, C. Gunter and M. Snyder (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Bienko, M., N. Crosetto, L. Teytelman, S. Klemm, S. Itzkovitz and A. van Oudenaarden (2013). "A versatile genome-scale PCR-based pipeline for high-definition DNA FISH." Nature Methods **10**(2): 122-124.

Bintu, B., L. J. Mateo, J. H. Su, N. A. Sinnott-Armstrong, M. Parker, S. Kinrot, . . . X. Zhuang (2018). "Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells." Science **362**(6413).

Boija, A., I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, . . . R. A. Young (2018). "Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains." Cell **175**(7): 1842-1855.e1816.

Bonev, B., N. Mendelson Cohen, Q. Szabo, L. Fritsch, G. L. Papadopoulos, Y. Lubling, . . . G. Cavalli (2017). "Multiscale 3D Genome Rewiring during Mouse Neural Development." Cell **171**(3): 557-572 e524.

Bouwman, B. A. and W. de Laat (2015). "Getting the genome in shape: the formation of loops, domains and compartments." Genome Biol **16**: 154.

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

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

Cairns, J., W. R. Orchard, V. Malysheva and M. Spivakov (2019). "Chicdiff: a computational pipeline for detecting differential chromosomal interactions in Capture Hi-C data." Bioinformatics **35**(22): 4764-4766.

Cardozo Gizzi, A. M., D. I. Cattoni, J. B. Fiche, S. M. Espinola, J. Gurgo, O. Messina, . . . M. Nollmann (2019). "Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms." Mol Cell **74**(1): 212-222 e215.

Cheatham, T. E., 3rd and D. A. Case (2013). "Twenty-five years of nucleic acid simulations." Biopolymers **99**(12): 969-977.

Chen, B., L. A. Gilbert, B. A. Cimini, J. Schnitzbauer, W. Zhang, G. W. Li, . . . B. Huang (2013). "Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system." Cell **155**(7): 1479-1491.

Chiariello, A. M., C. Annunziatella, S. Bianco, A. Esposito and M. Nicodemi (2016). "Polymer physics of chromosome large-scale 3D organisation." Sci Rep **6**: 29775.

Chiariello, A. M., S. Bianco, A. M. Oudelaar, A. Esposito, C. Annunziatella, L. Fiorillo, . . . M. Nicodemi (2020). "A Dynamic Folded Hairpin Conformation Is Associated with

α-Globin Activation in Erythroid Cells." Cell Rep **30**(7): 2125-2135.e2125.

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

Clapier, C. R., J. Iwasa, B. R. Cairns and C. L. Peterson (2017). "Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes." Nat Rev Mol Cell Biol **18**(7): 407-422.

Cremer, T. and M. Cremer (2010). "Chromosome territories." Cold Spring Harb Perspect Biol **2**(3): a003889.

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

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

Deal, R. B., J. G. Henikoff and S. Henikoff (2010). "Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones." Science **328**(5982): 1161-1164.

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

Di Stefano, M., A. Rosa, V. Belcastro, D. di Bernardo and C. Micheletti (2013). "Colocalization of coregulated genes: a steered molecular dynamics study of human chromosome 19." PLoS Comput Biol **9**(3): e1003019.

Dietzel, S., A. Jauch, D. Kienle, G. Qu, H. Holtgreve-Grez, R. Eils, . . . T. Cremer (1998). "Separate and variably shaped chromosome arm domains are disclosed by chromosome arm painting in human cell nuclei." Chromosome Res **6**(1): 25-33.

Dixon, J. R., I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, . . . B. Ren (2015). "Chromatin architecture reorganization during stem cell differentiation." Nature **518**(7539): 331-336.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, . . . B. Ren (2012). "Topological domains in mammalian genomes

identified by analysis of chromatin interactions." Nature **485**(7398): 376-380.

Djekidel, M. N., Y. Chen and M. Q. Zhang (2018). "FIND: difFerential chromatin INteractions Detection using a spatial Poisson process." Genome Res **28**(3): 412-422.

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

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

Fang, R., M. Yu, G. Li, S. Chee, T. Liu, A. D. Schmitt and B. Ren (2016). "Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq." Cell Research **26**(12): 1345-1348.

Filion, G. J., J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, . . . B. van Steensel (2010). "Systematic protein location mapping reveals five principal chromatin types in Drosophila cells." Cell **143**(2): 212-224.

Finch, J. T. and A. Klug (1976). "Solenoidal model for superstructure in chromatin." Proc Natl Acad Sci U S A **73**(6): 1897-1901.

Fraser, J., C. Ferrai, A. M. Chiariello, M. Schueler, T. Rito, G. Laudanno, . . . M. Nicodemi (2015). "Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation." Mol Syst Biol **11**(12): 852.

Fraser, P. and W. Bickmore (2007). "Nuclear organization of the genome and the potential for gene regulation." Nature **447**(7143): 413-417.

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

Ghavi-Helm, Y., A. Jankowski, S. Meiers, R. R. Viales, J. O. Korbel and E. E. M. Furlong (2019). "Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression." Nature genetics **51**(8): 1272-1282.

Goloborodko, A., J. F. Marko and L. A. Mirny (2016). "Chromosome Compaction by Active Loop Extrusion." Biophys J **110**(10): 2162-2168.

Greenwald, W. W., H. Li, P. Benaglio, D. Jakubosky, H. Matsui, A. Schmitt, . . . K. A. Frazer (2019). "Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression." Nature communications **10**(1): 1054-1054.

Grosberg, A., S. Nechaev and S. Ei (1988). "The role of topological constraints in the kinetics of collapse of macromolecules." Journal De Physique **49**: 2095-2100.

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

Gurumurthy, A., Y. Shen, E. M. Gunn and J. Bungert (2019). "Phase Separation and Transcription Regulation: Are Super-Enhancers and Locus Control Regions Primary Sites of Transcription Complex Assembly?" Bioessays **41**(1): e1800164.

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

Hsieh, T. S., C. Cattoglio, E. Slobodyanyuk, A. S. Hansen, O. J. Rando, R. Tjian and X. Darzacq (2020). "Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding." Mol Cell **78**(3): 539-553 e538.

Hsieh, T. S., G. Fudenberg, A. Goloborodko and O. J. Rando (2016). "Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome." Nat Methods **13**(12): 1009-1011.

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

Huang, P., C. A. Keller, B. Giardine, J. D. Grevet, J. O. J. Davies, J. R. Hughes, . . . G. A. Blobel (2017). "Comparative analysis of three-dimensional chromosomal architecture identifies a novel fetal hemoglobin regulatory element." Genes Dev **31**(16): 1704-1713.

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

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

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

Irastorza-Azcarate, I., R. D. Acemel, J. J. Tena, I. Maeso, J. L. Gomez-Skarmeta and D. P. Devos (2018). "4Cin: A computational pipeline for 3D genome modeling and virtual Hi-C analyses from 4C data." PLoS Comput Biol **14**(3): e1006030.

Jackson, D. A., A. B. Hassan, R. J. Errington and P. R. Cook (1993). "Visualization of focal sites of transcription within human nuclei." EMBO J **12**(3): 1059-1065.

Javierre, B. M., O. S. Burren, S. P. Wilder, R. Kreuzhuber, S. M. Hill, S. Sewitz, . . . P. Fraser (2016). "Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters." Cell **167**(5): 1369-1384 e1319.

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

Kapilevich, V., S. Seno, H. Matsuda and Y. Takenaka (2019). "Chromatin 3D Reconstruction from Chromosomal Contacts Using a Genetic Algorithm." IEEE/ACM Trans Comput Biol Bioinform **16**(5): 1620-1626.

Kempfer, R. and A. Pombo (2020). "Methods for mapping 3D chromosome architecture." Nat Rev Genet **21**(4): 207-226.

Kim, S., N.-K. Yu and B.-K. Kaang (2015). "CTCF as a multifunctional protein in genome regulation and gene expression." Experimental & Molecular Medicine **47**(6): e166-e166.

Kim, T. H., Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, . . . B. Ren (2007). "Analysis of the

vertebrate insulator protein CTCF-binding sites in the human genome." Cell **128**(6): 1231-1245.

Klemm, S. L., Z. Shipony and W. J. Greenleaf (2019). "Chromatin accessibility and the regulatory epigenome." Nat Rev Genet **20**(4): 207-220.

Knight, P. A. and D. Ruiz (2012). "A fast algorithm for matrix balancing." IMA Journal of Numerical Analysis **33**(3): 1029-1047.

Koslover, E. F., C. J. Fuller, A. F. Straight and A. J. Spakowitz (2010). "Local geometry and elasticity in compact chromatin structure." Biophysical journal **99**(12): 3941-3950.

Kremer, K. and G. S. Grest (1990). "Dynamics of entangled linear polymer melts: A molecular-dynamics simulation." The Journal of Chemical Physics **92**(8): 5057-5086.

Kullback, S. and R. A. Leibler (1951). "On Information and Sufficiency." Ann. Math. Statist. **22**(1): 79-86.

Kupper, K., A. Kolbl, D. Biener, S. Dittrich, J. von Hase, T. Thormeyer, . . . M. Cremer (2007). "Radial chromatin positioning is shaped by local gene density, not by gene expression." Chromosoma.

Lai, W. K. M. and B. F. Pugh (2017). "Understanding nucleosome dynamics and their links to gene expression and DNA replication." Nature Reviews Molecular Cell Biology **18**(9): 548-562.

Lakadamyali, M. and M. P. Cosma (2020). "Visualizing the genome in high resolution challenges our textbook understanding." Nat Methods **17**(4): 371-379.

Lee, J. H., E. R. Daugharthy, J. Scheiman, R. Kalhor, T. C. Ferrante, R. Terry, . . . G. M. Church (2015). "Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues." Nature Protocols **10**(3): 442-458.

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

Lin, D., G. Bonora, G. G. Yardımcı and W. S. Noble (2019). "Computational methods for analyzing and modeling genome structure and organization." Wiley Interdiscip Rev Syst Biol Med **11**(1): e1435.

Liu, X., Y. Zhang, Y. Chen, M. Li, F. Zhou, K. Li, . . . J. Xu (2017). "In Situ Capture of Chromatin Interactions by Biotinylated dCas9." Cell **170**(5): 1028-1043 e1019.

Luger, K., A. W. Mader, R. K. Richmond, D. F. Sargent and T. J. Richmond (1997). "Crystal structure of the nucleosome core particle at 2.8 A resolution." Nature **389**(6648): 251-260.

Lun, A. T. and G. K. Smyth (2015). "diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data." BMC Bioinformatics **16**: 258.

Lupianez, D. G., K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, . . . S. Mundlos (2015). "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions." Cell **161**(5): 1012-1025.

Maass, P. G., A. R. Barutcu and J. L. Rinn (2019). "Interchromosomal interactions: A genomic love story of kissing chromosomes." J Cell Biol **218**(1): 27-38.

Marti-Renom, M. A. and L. A. Mirny (2011). "Bridging the resolution gap in structural modeling of 3D genome organization." PLoS Comput Biol **7**(7): e1002125.

Mateo, L. J., S. E. Murphy, A. Hafner, I. S. Cinquini, C. A. Walker and A. N. Boettiger (2019). "Visualizing DNA folding and RNA in embryos at single-cell resolution." Nature.

Mifsud, B., I. Martincorena, E. Darbo, R. Sugar, S. Schoenfelder, P. Fraser and N. M. Luscombe (2017). "GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data." PLoS One **12**(4): e0174744.

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

Miguel-Escalada, I., S. Bonas-Guarch, I. Cebola, J. Ponsa-Cobas, J. Mendieta-Esteban, G. Atla, . . . J. Ferrer (2019). "Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes." Nat Genet **51**(7): 1137-1148.

Mirny, L. A. (2011). "The fractal globule as a model of chromatin architecture in the cell." Chromosome Res **19**(1): 37-51.

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

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

Nagano, T., Y. Lubling, E. Yaffe, S. W. Wingett, W. Dean, A. Tanay and P. Fraser (2015). "Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell." Nat Protoc **10**(12): 1986-2003.

Nguyen, H. Q., S. Chattoraj, D. Castillo, S. C. Nguyen, G. Nir, A. Lioutas, . . . C. T. Wu (2020). "3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing." Nat Methods **17**(8): 822-832.

Nir, G., I. Farabella, C. Perez Estrada, C. G. Ebeling, B. J. Beliveau, H. M. Sasaki, . . . C. T. Wu (2018). "Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling." PLoS Genet **14**(12): e1007872.

Nora, E. P., A. Goloborodko, A.-L. Valton, J. H. Gibcus, A. Uebersohn, N. Abdennur, . . . B. Bruneau (2017). "Targeted degradation of CTCF decouples local insulation of chromosome domains from higher-order genomic compartmentalization." bioRxiv.

Nora, E. P., A. Goloborodko, A. L. Valton, J. H. Gibcus, A. Uebersohn, N. Abdennur, . . . B. G. Bruneau (2017). "Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization." Cell **169**(5): 930-944 e922.

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

Norton, H. K. and J. E. Phillips-Cremins (2017). "Crossed wires: 3D genome misfolding in human disease." J Cell Biol **216**(11): 3441-3452.

Oluwadare, O., M. Highsmith and J. Cheng (2019). "An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data." Biological Procedures Online **21**(1): 7.

Osborne, C. S., L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, . . . P. Fraser (2004). "Active genes dynamically

colocalize to shared sites of ongoing transcription." Nat Genet **36**(10): 1065-1071.

Osborne, C. S., L. Chakalova, J. A. Mitchell, A. Horton, A. L. Wood, D. J. Bolland, . . . P. Fraser (2007). "Myc dynamically and preferentially relocates to a transcription factory occupied by Igh." PLoS Biol **5**(8): e192.

Ou, H. D., S. Phan, T. J. Deerinck, A. Thor, M. H. Ellisman and C. C. O'Shea (2017). "ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells." Science **357**(6349).

Oudelaar, A. M., J. O. J. Davies, L. L. P. Hanssen, J. M. Telenius, R. Schwessinger, Y. Liu, . . . J. R. Hughes (2018). "Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains." Nat Genet **50**(12): 1744-1751.

Pachov, G. V., R. R. Gabdoulline and R. C. Wade (2011). "On the structure and dynamics of the complex of the nucleosome and the linker histone." Nucleic acids research **39**(12): 5255-5263.

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

Pierce, B. A. (2012). "Genetics: A conceptual approach." New York : W.H. Freeman/Macmillan Learning **2nd edition**.

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

Pott, S. and J. D. Lieb (2015). "What are super-enhancers?" Nature Genetics **47**(1): 8-12.

Pueschel, R., F. Coraggio and P. Meister (2016). "From single genes to entire genomes: the search for a function of nuclear organization." Development **143**(6): 910-923.

Quinodoz, S. A., P. Bhat, N. Ollikainen, J. W. Jachowicz, A. K. Banerjee, P. Chovanec, . . . M. Guttman (2020). "RNA promotes the formation of spatial compartments in the nucleus." bioRxiv: 2020.2008.2025.267435.

Quinodoz, S. A., N. Ollikainen, B. Tabak, A. Palla, J. M. Schmidt, E. Detmar, . . . M. Guttman (2018). "Higher-Order Inter-

chromosomal Hubs Shape 3D Genome Organization in the Nucleus." Cell **174**(3): 744-757 e724.

Ramani, V., X. Deng, R. Qiu, C. Lee, C. M. Disteche, W. S. Noble, . . . Z. Duan (2020). "Sci-Hi-C: A single-cell Hi-C method for mapping 3D genome organization in large number of single cells." Methods **170**: 61-68.

Rao, S. S., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, . . . E. L. Aiden (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." Cell **159**(7): 1665-1680.

Rao, S. S. P., S. C. Huang, B. Glenn St Hilaire, J. M. Engreitz, E. M. Perez, K. R. Kieffer-Kwon, . . . E. L. Aiden (2017). "Cohesin Loss Eliminates All Loop Domains." Cell **171**(2): 305-320 e324.

Ricci, M. A., C. Manzo, M. F. Garcia-Parajo, M. Lakadamyali and M. P. Cosma (2015). "Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo." Cell **160**(6): 1145-1158.

Rieber, L. and S. Mahony (2017). "miniMDS: 3D structural inference from high-resolution Hi-C data." Bioinformatics **33**(14): i261-i266.

Robson, M. I., A. R. Ringel and S. Mundlos (2019). "Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D." Mol Cell **74**(6): 1110-1122.

Rosa, A. and R. Everaers (2008). "Structure and dynamics of interphase chromosomes." PLoS Comput Biol **4**(8): e1000153.

Russel, D., K. Lasker, B. Webb, J. Velazquez-Muriel, E. Tjioe, D. Schneidman-Duhovny, . . . A. Sali (2012). "Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies." PLoS Biol **10**(1): e1001244.

Sanyal, A., D. Bau, M. A. Marti-Renom and J. Dekker (2011). "Chromatin globules: a common motif of higher order chromosome structure?" Curr Opin Cell Biol **23**(3): 325-331.

Schmitt, A. D., M. Hu, I. Jung, Z. Xu, Y. Qiu, C. L. Tan, . . . B. Ren (2016). "A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome." Cell Rep **17**(8): 2042-2059.

Schneider, R. and R. Grosschedl (2007). "Dynamics and interplay of nuclear architecture, genome organization, and gene expression." Genes Dev **21**(23): 3027-3043.

Schoenfelder, S., B. M. Javierre, M. Furlan-Magaril, S. W. Wingett and P. Fraser (2018). "Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions." J Vis Exp(136).

Schones, D. E., K. Cui, S. Cuddapah, T. Y. Roh, A. Barski, Z. Wang, . . . K. Zhao (2008). "Dynamic regulation of nucleosome positioning in the human genome." Cell **132**(5): 887-898.

Schübeler, D., C. Francastel, D. M. Cimbora, A. Reik, D. I. Martin and M. Groudine (2000). "Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human beta-globin locus." Genes & development **14**(8): 940-950.

Sedeno Cacciatore, A. and B. D. Rowland (2019). "Loop formation by SMC complexes: turning heads, bending elbows, and fixed anchors." Curr Opin Genet Dev **55**: 11-18.

Serra, F., D. Bau, M. Goodstadt, D. Castillo, G. J. Filion and M. A. Marti-Renom (2017). "Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors." PLoS Comput Biol **13**(7): e1005665.

Serra, F., M. Di Stefano, Y. G. Spill, Y. Cuartero, M. Goodstadt, D. Bau and M. A. Marti-Renom (2015). "Restraint-based three-dimensional modeling of genomes and genomic domains." FEBS Lett **589**(20 Pt A): 2987-2995.

Servant, N., N. Varoquaux, B. R. Lajoie, E. Viara, C. J. Chen, J. P. Vert, . . . E. Barillot (2015). "HiC-Pro: an optimized and flexible pipeline for Hi-C data processing." Genome Biol **16**: 259.

Shachar, S., G. Pegoraro and T. Misteli (2015). "HIPMap: A High-Throughput Imaging Method for Mapping Spatial Gene Positions." Cold Spring Harb Symp Quant Biol.

Siggens, L. and K. Ekwall (2014). "Epigenetics, chromatin and genome organization: recent advances from the ENCODE project." J Intern Med **276**(3): 201-214.

Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, . . . W. de Laat (2006). "Nuclear organization of active and inactive chromatin domains uncovered by

chromosome conformation capture-on-chip (4C)." Nat Genet **38**(11): 1348-1354.
Soutourina, J. (2018). "Transcription regulation by the Mediator complex." Nat Rev Mol Cell Biol **19**(4): 262-274.
Stik, G., E. Vidal, M. Barrero, S. Cuartero, M. Vila-Casadesús, J. Mendieta-Esteban, . . . T. Graf (2020). "CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response." Nature Genetics **52**(7): 655-661.
Su, J. H., P. Zheng, S. S. Kinrot, B. Bintu and X. Zhuang (2020). "Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin." Cell.
Sutherland, H. and W. A. Bickmore (2009). "Transcription factories: gene expression in unions?" Nat Rev Genet **10**(7): 457-466.
Szabo, Q., F. Bantignies and G. Cavalli (2019). "Principles of genome folding into topologically associating domains." Science Advances **5**(4): eaaw1668.
Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, . . . J. A. Stamatoyannopoulos (2012). "The accessible chromatin landscape of the human genome." Nature **489**(7414): 75-82.
Trieu, T. and J. Cheng (2017). "3D genome structure modeling by Lorentzian objective function." Nucleic Acids Res **45**(3): 1049-1058.
Trieu, T., O. Oluwadare and J. Cheng (2019). "Hierarchical Reconstruction of High-Resolution 3D Models of Large Chromosomes." Scientific Reports **9**(1): 4971.
Trussart, M., F. Serra, D. Bau, I. Junier, L. Serrano and M. A. Marti-Renom (2015). "Assessing the limits of restraint-based 3D modeling of genomes and genomic domains." Nucleic Acids Res **43**(7): 3465-3477.
Ulianov, S. V., A. A. Gavrilov and S. V. Razin (2015). "Nuclear compartments, genome folding, and enhancer-promoter communication." Int Rev Cell Mol Biol **315**: 183-244.
Ulianov, S. V., E. E. Khrameeva, A. A. Gavrilov, I. M. Flyamer, P. Kos, E. A. Mikhaleva, . . . S. V. Razin (2016). "Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains." Genome Res **26**(1): 70-84.
van de Werken, H. J., P. J. de Vree, E. Splinter, S. J. Holwerda, P. Klous, E. de Wit and W. de Laat (2012). "4C technology:

protocols and data analysis." Methods Enzymol **513**: 89-112.

van Steensel, B. and E. E. M. Furlong (2019). "The role of transcription in shaping the spatial organization of the genome." Nat Rev Mol Cell Biol **20**(6): 327-337.

Vermeulen, C., A. Allahyar, B. A. M. Bouwman, P. H. L. Krijger, M. J. A. M. Verstegen, G. Geeven, . . . W. de Laat (2020). "Multi-contact 4C: long-molecule sequencing of complex proximity ligation products to uncover local cooperative and competitive chromatin topologies." Nature Protocols **15**(2): 364-397.

Vidal, E., F. le Dily, J. Quilez, R. Stadhouders, Y. Cuartero, T. Graf, . . . G. J. Filion (2018). "OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes." Nucleic Acids Res **46**(8): e49.

Vilarrasa-Blasi, R., P. Soler-Vila, N. Verdaguer-Dot, N. Russiñol, M. Di Stefano, V. Chapaprieta, . . . J. I. Martin-Subero (2019). "Dynamics of genome architecture and chromatin function during human B cell differentiation and neoplastic transformation." bioRxiv: 764910.

Volpi, E. V. and J. M. Bridger (2008). "FISH glossary: an overview of the fluorescence in situ hybridization technique." Biotechniques **45**(4): 385-386, 388, 390 passim.

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

Wani, A. H., A. N. Boettiger, P. Schorderet, A. Ergun, C. Munger, R. I. Sadreyev, . . . N. J. Francis (2016). "Chromatin topology is coupled to Polycomb group protein subnuclear organization." Nat Commun **7**: 10291.

Weintraub, A. S., C. H. Li, A. V. Zamudio, A. A. Sigova, N. M. Hannett, D. S. Day, . . . R. A. Young (2017). "YY1 Is a Structural Regulator of Enhancer-Promoter Loops." Cell **171**(7): 1573-1588.e1528.

Wingett, S., P. Ewels, M. Furlan-Magaril, T. Nagano, S. Schoenfelder, P. Fraser and S. Andrews (2015). "HiCUP: pipeline for mapping and processing Hi-C data." F1000Res **4**: 1310.

Xie, S. Q., L. M. Lavitas and A. Pombo (2010). "CryoFISH: fluorescence in situ hybridization on ultrathin cryosections." Methods Mol Biol **659**: 219-230.

Zabidi, M. A. and A. Stark (2016). "Regulatory Enhancer-Core-Promoter Communication via Transcription Factors and Cofactors." Trends Genet **32**(12): 801-814.

Zhang, P., W. Wu, Q. Chen and M. Chen (2019). "Non-Coding RNAs and their Integrated Networks." Journal of integrative bioinformatics **16**(3): 20190027.

Zhu, G., W. Deng, H. Hu, R. Ma, S. Zhang, J. Yang, . . . J. Zeng (2018). "Reconstructing spatial organizations of chromosomes through manifold learning." Nucleic Acids Res **46**(8): e50.

Zhu, H. and Z. Wang (2019). "SCL: a lattice-based approach to infer 3D chromosome structures from single-cell Hi-C data." Bioinformatics **35**(20): 3981-3988.