Three-dimensional regulation: Establishing novel linkages between non-coding genetic variation and target genes

Ning Liu

Submitted for the degree of Doctor of Philosophy

Discipline of Paediatrics

Medical School

The University of Adelaide

April 2021

Table of Contents

List of included publications

1. **Liu, N.**, Low, W. Y., Alinejad-Rokny, H., Pederson, S., Sadlon, T., Barry, S., & Breen, J. (2021). Seeing the forest through the trees: prioritising potentially functional interactions from Hi-C. Epigenetics & Chromatin, 14(1), 1-17.

2. **Liu, N.**, Sadlon, T., Wong, Y. Y., Pederson, S. M., Breen, J., & Barry, S. C. (2020). 3DFAACTS-SNP: Using regulatory T cell-specific epigenomics data to uncover candidate mechanisms of Type-1 Diabetes (T1D) risk. bioRxiv.

3. **Liu, N.**, Alinejad-Rokny, H., Breen, J. (2021). Landscape of statistically significant chromatin interaction profiles of cell lines and primary tissues in the human genome.

4. **Liu, N.**, Alinejad-Rokny, H., Breen, J. (2021). Identifying human cell/tissue type-specific potentially functional interactions.

Abstract

In the human genome, 98% of the DNA is in non-gene coding regions. While these regions do not express genes, a mounting number of studies have shown that they are crucial to the maintenance of chromosome structure and in the regulation of gene expression. Although large epigenomics projects were established to functionally annotate non-coding regions, the comprehensive linkages between these regions and their target genes remain unknown. The human genome folds into hierarchical three-dimensional (3D) structure, bringing distantly regulatory elements into close proximity, leading to the formation of 3D chromatin physical interactions and playing an important role in the complex gene regulation network. Using chromatin interaction information, we can connect functional non-coding regions to their target genes to reveal novel regulation mechanisms.

In Chapter 1, we reviewed current existing approaches to prioritise functional interactions from Hi-C data, the state-of-the-art data type used to study chromatin interactions, and categorised them into three classifications, including structural-based methods, statistical model-based methods and data integration methods. Chapter 2 described the computational procedures of analysing Hi-C datasets, and introduced: HiC-QC, a tool that extracting summary statistics to perform quality control with Hi-C libraires; *HiC-interactionmap* and *integration-tracks* plot, tools to offer visualisation for Hi-C data integration. Additionally, aligners BWA and Bowtie2, were compared for their performance of mapping Hi-C data.

Using type 1 diabetes (T1D) and regulatory T cells (Treg) as a disease-cell type model, based on data integration of Treg-specific Hi-C interactions and other epigenomics information, Chapter 3 established a filtering workflow called 3DFAACT-SNPs to link genetic variants that are associated with T1D to the loss of immune tolerance in Treg. Using this workflow, we identified 36 SNPs with plausible Treg-specific mechanisms of action contributing to T1D, linking 119 novel interacting regions. We demonstrated that it is possible to prioritise SNPs that contribute to disease based on regulatory function and illustrate the power of using chromatin interactions to connect non-coding SNPs to disease mechanisms.

Lastly, Chapters 4 and 5 launch the statistically significant interaction profiling of 51 human cell lines and primary tissues from 173 public Hi-C datasets using a statistical model from MaxHiC, followed by investigating the uniqueness, distancing preference and the associated genes of the cell/tissue-specific interactions. We also identified interaction “hot zones”, regions with chromatin interactions observed across many cells and tissues. Using global and local enrichment analysis and a comparison to frequent interacting regions, we demonstrated the structural and regulatory functionality of the hot zones. We further comprehensively annotated chromatin interactions into 66 interaction classes, cataloguing potentially regulatory functional interactions for different cells and tissues. Finally, we revealed cell/tissue-specific 3D regulatory regions that are enriched with super-enhancers and overlapped with expression quantitative trait loci (eQTLs).

Overall, using data integration and statistical models to prioritise functional chromatin interactions, this work produced novel computational tools and pipelines and generated valuable resource for the investigation of genome structure, demonstrating the power of using chromatin interactions to discover novel mechanisms in the genome and revealing novel linkages between non-coding DNA to traits/diseases.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Ning Liu Date: 21 April 2021

Acknowledgements

This dissertation becomes a reality with the generous support and help from many individuals. I am gratefully indebted to them and would like to express my sincere thanks to all of them.

First and foremost, I would like to express my utmost gratitude to my primary supervisor **Jimmy Breen** for his incredible and unwavering patience in mentoring and supervising me throughout my research over the past four years. I never regretted a single moment of being your student to establish my foundation to become a researcher. Your vision, motivation, integrity, wisdom and humour have greatly inspired me and shaped me into a better person and researcher. I also would like to extend my thanks to your family, Nadine and Frank, for their understanding during the preparation of this thesis.

Secondly, I would like to thank my co-supervisors **Simon Barry** and **Rick Tearle**, and my collaborators and colleges, including **Hamid Alinejad-Rokny**, **Stephen Pederson**, **Wai Yee Low**, **Timothy Sadlon**, **Melaine Smith** and **Jacqueline Rehn**. Thank you all for all your extremely helpful suggestions during the preparation of this thesis. And thank you for sharing your insights and knowledge in all the meetings we had during the last couple years.

Thirdly, I would like to thank all members and friends of the Bioinformatics Hub. Specifically **Alistair Ludington**, **Nhi Hin**, **Qianhui Wan**, **Justin Bogias**,

Lachlan Baer, **Chelsea Matthews, Yingying Wong**, **Dan Kortschak**, **Nathan Watson-Haigh**, **Terry Bertozzi** and **David Adelson**. Thank you all for making the Bioinformatics Hub such a friendly environment, I would never forget the time in the hub when we had our weekly meetings of sharing cool bioinformatics tools, research ideas and outcomes and criticizing research papers.

Lastly, I would like to express my heartfelt gratitude for the support of my family. Particularly, my parents, for their unconditional love, and for all the financial support during my PhD, especially during the COVID-19 restriction; and my partner **Shuwen Cen**, for always believing in me, and for all the understanding, tolerance and moral support.

Chapter 1

Seeing the forest through the trees:

Identifying potentially functional interactions

from Hi-C

Seeing the forest through the trees: Prioritising potentially functional interactions from Hi-C

Ning Liu[1,2,3*], Wai Yee Low[4], Hamid Alinejad-Rokny[5,6], Stephen Pederson[3,9], Timothy Sadlon[2,7], Simon Barry[276], James Breen[1,2,3,8*]

[1] South Australian Health & Medical Research Institute, Adelaide, Australia

[2] Robinson Research Institute, University of Adelaide, Adelaide, Australia

[3] Adelaide Medical School, University of Adelaide, Adelaide, Australia

[4] The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, SA, 5371, Australia

[5] BioMedical Machine Learning Lab, The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, NSW, 2052, Australia

[6] Core Member of UNSW Data Science Hub, The University of New South Wales (UNSW Sydney), Sydney, 2052, Australia.

[7] Women's & Children's Health Network, Adelaide, Australia

[8] South Australian Genomics Centre (SAGC), Adelaide, Australia

[9] Dame Roma Mitchell Cancer Research Laboratories (DRMCRL), Adelaide Medical School, University of Adelaide, Australia

* Corresponding authors: Ning Liu (ning.liu@adelaide.edu.au) & James Breen (jimmy.breen@sahmri.com)

Abstract

Eukaryotic genomes are highly organised within the nucleus of a cell, allowing widely dispersed regulatory elements such as enhancers to interact with gene promoters through physical contacts in three-dimensional space. Recent chromosome conformation capture methodologies such as Hi-C have enabled the analysis of interacting regions of the genome providing a valuable insight into the three-dimensional organisation of the chromatin in the nucleus, including chromosome compartmentalisation and gene expression. Complicating the analysis of Hi-C data however is the massive amount of identified interactions, many of which do not directly drive gene function, thus hindering the identification of potentially biologically functional 3D interactions. In this review, we collate and examine the downstream analysis of Hi-C data with particular focus on methods that prioritise potentially functional interactions. We classify three groups of approaches; structural-based discovery methods e.g. A/B compartments and topologically-associated domains, detection of statistically significant chromatin interactions, and the use of epigenomic data integration to narrow down useful interaction information. Careful use of these three approaches is crucial to successfully identifying potentially functional interactions within the genome.

Keywords

chromosome conformation capture, Hi-C, statistically significant interactions identification, data integration

Background

The three-dimensional (3D) architecture of the eukaryotic genome has been shown to be an important factor in regulating transcription [1–3]. In the nucleus, DNA is folded into a highly organised structure, allowing transcriptional and regulatory machinery to be in specific nuclear territories for efficient usage. The impact of DNA folding and the resulting physical interactions can have dramatic impacts on the regulation of the genes, enabling non-coding regions such as regulatory elements (e.g. enhancers and silencers) to act on distally located gene promoters with disruption of chromosomal organisation increasingly linked to disease [4–6]. However, while highly organised, the folding structure of the 3D genome can also be highly dynamic to allow for the flexibility and modularity to facilitate regulatory action across a wide-range of cell-types and biological processes, such as development, immune homeostasis, cancer and diseases.

In recent decades, the development of chromosome conformation capture assays and high-throughput sequencing has facilitated the construction of 3D genomes at high resolution, enabling the identification of cell-type and tissue-specific 3D interactions between regions in the genome. However, the analysis of such data is complicated by the massive amount of identified physical interactions, hindering the detection and interpretation of interactions that are biologically meaningful. In this review, we introduce the background of 3D genome structure and its components, followed by a summary of the protocols that are commonly used to study 3D genome architecture in recent years, focusing on Hi-C protocols and other derived methods, whilst the use of microscopy to image 3D genome organization has also been recently reviewed [7]. We then thoroughly review current *in silico* methods for identification of potentially functional interactions, which are contacts with higher chance to be biologically functionally-relevant, and categorise them into three methodological groups.

Chromosome architecture and gene regulation

Within eukaryotic nuclei, chromosomal DNA is condensed and folded into highly organised 3D structures, with distinct functional domains [8,9]. A key consequence of chromosome folding is that it can bring DNA regions that are far away from each other on the same linear DNA polymer (i.e. intrachromosomal), into close proximity, allowing direct physical contact to be established between regions. Interchromosomal interactions may also play an important role in transcriptional regulation but are less studied. The best characterised examples of this type of interaction include the clustering of ribosomal genes to form the nucleolus and the clustering of olfactory receptor genes to ensure the monogenic and mono-allelic expression in an individual olfactory neuron [10].

The most basic level of chromosome organisation is chromatin “Loop” structures (Figure 1A). Chromatin loops are formed based on a loop extrusion model, where linear DNA is squeezed out through the structural maintenance of chromosomes (SMC) cohesin complex until the complex encounters convergent CTCF bound at loop anchor sequences [8,11–14]. Chromatin loops can either bring distal enhancers and gene promoters into close proximity to increase gene expression, or exclude an enhancer away from the loop to initiate boundaries to repress gene expression [15–17]. The archetypal chromatin looping factors are the CCCTC-binding protein (CTCF) and Cohesin complex [18–20], with the initial transient chromatin loops are created by the Cohesin complex during the extrusion process, or anchored on one CTCF binding site while the other anchor moving dynamically [11,21,22]. Moreover, specific transcription factors such as EKLF, GATA-1, FOG-1, NANOG and YY1 [23–28] were confirmed to play important roles in the regulation of chromatin looping.

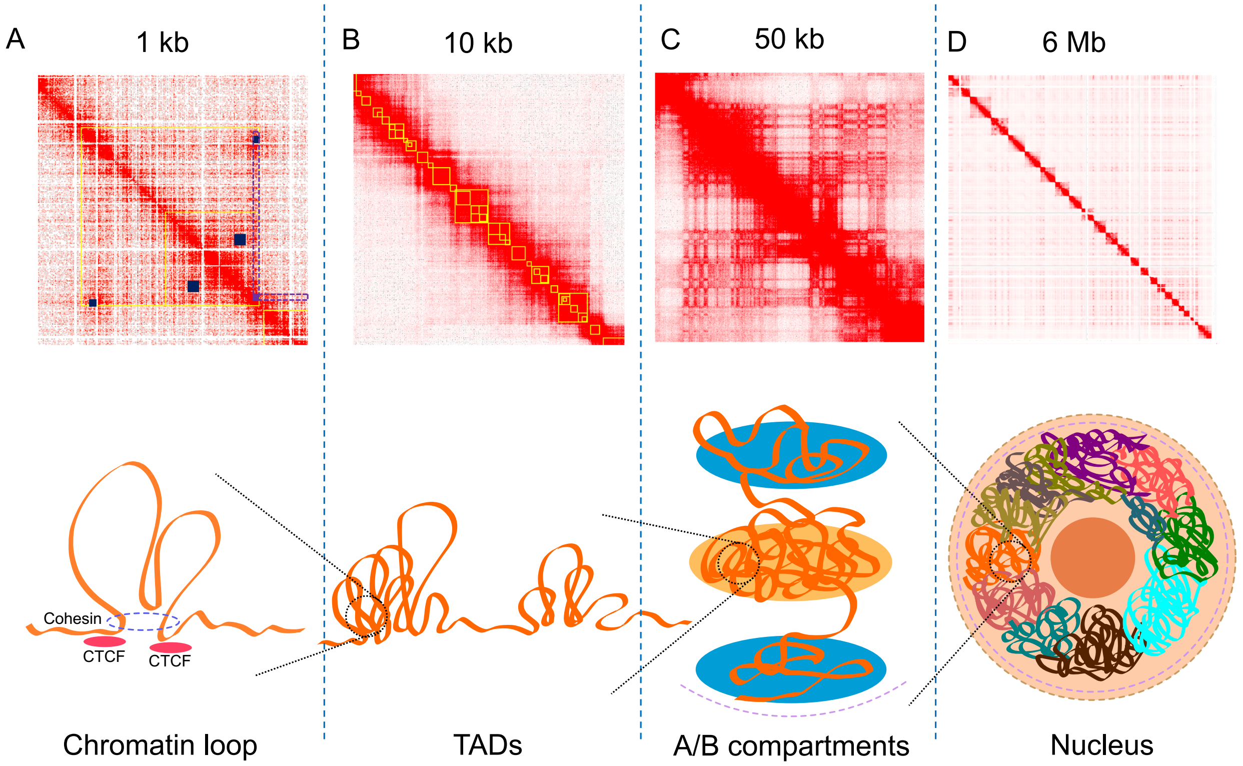


Figure 1: Illustration of genome architecture and the corresponding Hi-C interaction maps. Top panel: interaction heatmaps A, B, C, D are in different scales (kb or Mb per pixel) to correlate with the diagrams of 3D structures in the bottom panel, yellow boxes in A and B are identified TADs and small blue boxes in A indicate chromatin loops. The purple box in A is a frequently interacting region, with its classical "V" shape pattern colored in purple dotted lines. Heatmaps were generated using Juicebox [29] with published Hi-C data of GM12878 [3]. Bottom panel: diagrams of 3D structures in the genome.

Chromatin folding and DNA looping in particular leads to the formation of large scale chromatin structures such as topologically-associated domains (TADs) and chromosome compartments (Figure 1B) [30]. TADs are defined by chromatin interactions occurring more frequently within the TAD boundaries, with TAD boundaries often demarcating a change in interaction frequency [30]. TAD boundaries are also enriched for the insulator-binding protein CTCF and cohesin complex [19,20]. CTCF motif orientation appears to play a role in demarking TAD boundaries with some studies indicating that the majority of identified TADs (~60-

90%) have a CTCF motif at both anchor boundaries with convergent orientation [3,31,32]. This is consistent with the loop extrusion model mentioned above, suggesting that the formation of most TADs are form by extrusion and are strictly confined by boundaries established by ‘architectural’ proteins such as CTCF and SMC cohesin complex [33], along with the boundaries engaging with strong 3D interactions [34]. Moreover, experimental inversion of CTCF orientation or complete removal of the CTCF binding sites have been shown to disrupt the formation or shift the boundary of a TAD [14,16,32], further emphasizing the important role of CTCF defining TAD boundaries. The size of TADs are highly dependent on the resolution of the data and the chosen TAD caller and parameters [35], it can vary from hundreds of kilobases (kb) to 5 megabases (Mb) in mammalian genomes [36,37], and also show significant conservation in related species [38], suggesting that they may serve as the functional base of genome structure and development. With higher sequencing depth, patterns of interactions across regions within a TAD can be further divided into “sub-TADs” with a median size of 185 kb using one kilobase resolution data [3], enabling finer scale investigation of the genome structure [39,40]. In addition to “sub-TADs”, many other terms of TADs with different sizes and features have been proposed, including “micro-TADs” [41], “mega-domains” [42] and “super-TADs” [43]. However, functional distinction between the “conventional TADs” and them is still unclear. Evidence has shown that TADs are crucial structural units of long-range gene regulation [44–47], with interactions such as promoter-enhancer looping mostly found within the same TADs [48], and abnormal interactions across TADs (inter-TADs) can lead to significant regulation of expression level of important genes [49].

At a multi-megabase scale, the genome organisation is spatially segregated into euchromatin (gene-rich regions) or heterochromatin (gene-poor regions) to form active and inactive domains called ‘Compartments’ (Figure 1C) [2]. This compartmentalisation of chromosome folding depicts the global organisation of chromosomes in the nucleus, where compartment A corresponds to gene-dense, euchromatic regions, and compartment B corresponding to gene-poor heterochromatin. Using higher resolution data, the genome can be further grouped into six sub-compartments, compartment A is separated into A1 and A2 whereas compartment B is separated into B1, B2, B3 and B4, with each one associated with specific histone marks [3]. Sub-compartments A1 and A2 are enriched with active genes and the activating histone marks H3K4me3, H3K36me3, H3K27ac and H3K4me1. Sub-compartments A1 and A2 are also depleted in nuclear lamina and nucleolus-associated domains (NADs). B1 domains correlate with H3K27m[3]e3 positively and H3K36me3 negatively, B2 and B3 are enriched in nuclear lamina but B3 is depleted in NADs, and B4 is a 11 Mb region, containing lots of KRAB-ZNF genes [3].

The interaction of transcription factors bound at regulatory elements, such as promoters, enhancers and super-enhancers, mediate the transcription level of a gene via interactions which are the direct result of the 3D chromosome structure, but which appear to be long-distance interactions when viewed through lens of a linear chromosome [50–52]. One early and well-characterized example is the interaction between beta-globin locus and its locus control region (LCR) [53]. During the development and differentiation of erythroid in human and mouse, the LCR, which is located 40-60 kb away from beta-globin genes, contains the hypersensitive sites that are exhibiting strong enhancer function and contacting to beta-globin genes distally via chromatin loops to regulation gene expressions [54–56]. *Hox* gene clusters, essential for patterning the vertebrate body axis, are also governed by a rich enhancer interaction network. Using chromatin conformation capture methods, a number of studies found that the transcriptional activation or inactivation of *Hox* clusters requires a bimodal transition between active and inactive chromatin [30,57–60]. Taken together, the 3D genome structure governing long-distance contacts can build complex gene regulatory networks, allowing for either multiple enhancers to interact with a single promoter or a single enhancer to contact multiple promoters [61]. Disruption of these long range regulatory networks is increasingly being linked to both monogenic and complex diseases [62,63].

Hi-C assays to quantify chromatin interactions

In order to investigate the 3D genome architecture, a series of protocols called chromosome conformation capture (3C) assays have been developed that specifically capture the physical interactions between regions of DNA [1,2,64–

66]. A suite of 3C-derived high throughput DNA sequencing assays have been developed, including circular chromosome conformation capture sequencing (4C-seq) [64,67], chromosome conformation capture carbon copy (5C) [65], chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [66], enrichment of ligation products (ELP) [68] and higher-resolution chromosome conformation capture sequencing (Hi-C) [2], which vary in complexity or the scale of the interactions that are captured. The initial 3C method used PCR to quantify specific ligation products between a target sequence and a small number of defined regions [1]. 4C-seq, known as the "one vs all" method, uses an inverse PCR approach to convert all chimeric molecules associated with a specific region of interest generated in the proximity ligation step into a high throughput DNA sequencing library [67]. 5C increased the number of regions that could be captured by multiplexing PCR reactions [65], and it is also considered as the first "many vs many" approach and has been used to examine the long range interactions of between transcription start sites and approximately 1% of the human genome [69]. ChIA-PET implements a similar approach, however uses a specific, bound protein, generally a transcription factor protein, generating a protein-centric interaction profile [31]. ELP implements a double digestion strategy to improve the enrichment of 3C products in the library and is able to generate a detailed genome-wide contact map of the yeast genome [68].

Compared to other approaches, Hi-C, also known as the Genome conformation capture method [70], is the first "all vs all" method of genome-wide, 3C-derived assay to capture all interactions in the nucleus, allowing for a more complete snapshot of nuclear conformation at the global level [36]. Hi-C works through cross-linking DNA molecules in close proximity via a formaldehyde treatment, preserving the 3D interaction between two genomic regions. The cross-linked DNA is then usually fragmented using a restriction enzyme, such as the 6bp recognition enzyme *HindIII* [30,71] or 4bp cutter *MboI*, *DpnII* and *Sau3AI*, and the resultant DNA, ends held in close spatial proximity by the DNA cross-links, are ligated into chimeric DNA fragments. Subsequent steps convert these chimeric DNA fragments into linear fragments to which sequencing adapters are added to create a Hi-C library. The library is then sequenced using high-throughput sequencing technology, specifically limited to Illumina paired-end (as opposed to single-end/fragment) DNA sequencing to enable the accurate identification of the two ends of the hybrid molecule [2]. In the initial development of Hi-C, the identification of Hi-C interactions was impacted by the number of spurious ligation products generated as a result of the ligation step being carried in solution allowing for greater freedom for random inter-complex ligation reactions to occur. The resolution of Hi-C interactions in these earlier approaches was also limited by the cutting frequency of a 6-base restriction enzyme, such as *HindIII [2,30,72–74]*. To address these issues, an *in situ* Hi-C protocol was developed [3], where the ligation steps were performed within the constrained space of the nuclei, reducing the chance of random ligation [75,76]. Furthermore, *in situ* Hi-C used a 4-base-cutter (such as *MboI*) for digestion, increasing the cutting frequency in the genome and improving the resolution of captured interactions

[3]. Using this method, the first 3D map of the human genome was constructed using the GM12878 cell line with approximately 4.9 billion interactions [3], enabling interaction resolution at the kilobase level. In recent years, the *in situ* Hi-C protocol has been developed further to target different technical and/or biological questions (Table 1).

Table 1: Different Hi-C-derived methods. Optimizations indicate their modification in their protocols compared to traditional Hi-C.

Hi-C flavours	Optimizations	Advantages compared to traditional Hi-C	Reference
Traditional Hi-C	-	-	[2]
in situ Hi-C	Nuclear ligation; 4-based cutter	Allow higher resolution data generation	[3]
DNase Hi-C	DNase I to digest crosslinked DNA	Improve capture efficiency, reducing digestion bias but have A compartment bias	[77]
Micro-C	Crosslinking with DSG and micrococcal nuclease to digest crosslinked DNA	Improve capture efficiency, reducing digestion bias but have A compartment bias	[78]
BL-Hi-C	HaeIII to digest crosslinked DNA, followed by a two-step ligation	Improve capture efficiency in regulatory regions, reducing random ligation events	[79]
DLO Hi-C	No labelling and pull-down step	Reduce experimental cost	[80]
tag Hi-C	Tn5-transposase tagmentation	Focus on accessible chromatin, allow only hundreds of cells as input, reduce experimental cost	[81]
Capture HiC	RNA baits to subset specific chromatin contacts	Reduce sequencing cost, focus on a subset of interactions	[82]
Capture-C/NG Capture-C/Tiled-C	Enrich the 3C library with biotinylated capture oligonucleotides	Focus on the subset of interactions while retaining maximal library complexity	[83], [84], [85]
HiChIP/PLAC-seq	Chromatin Immunoprecipitation (ChIP) to subset bound chromatin contacts	Reduce sequencing cost, focus on a subset of interactions	[86], [87]

OCEAN-C	Phenol-chloroform extraction step	Focus on accessible chromatin	[88]
HiCoP	Column purified chromatin step	Focus on accessible chromatin	[89]
Methyl-HiC	Bisulfite conversion	Allow jointly profiling of DNA methylation and 3D genome structure	[90]
Hi-C 2.0	Efficient unligated ends removal	Largely reduce the dangling end DNA products	[91]
Hi-C 3.0	Double crosslinking with FA and DSG and double digestion with *DpnII* and *DdeI*	Improve the ability to identify A/B compartments and improve the enrichment of regulatory elements in loop detection	[92]

Owing to the vast complexity of the Hi-C ligation products generated, it is often too costly to sequence samples to a sufficient depth to achieve the resolution necessary to investigate specific interactions such as promoter-enhancer interactions, leading to the development of capture Hi-C (CHi-C) [82]. CHi-C employs a sequence capture approach, using pools of probes complementary to thousands of restriction fragments, to enrich for molecules containing the region of interest from the Hi-C library. This significantly reduces the complexity of the libraries and enables a significant increase in the number of detectable interactions within specific regions without the need for ultra-deep sequencing. Therefore CHi-C, has been used in many cases to analyse specific types of long-range interactions, such as interactions linked to promoter or enhancer regions. For example, CHi-C was recently used to characterise promoter interactions in 17 human primary hematopoietic cells to demonstrate the highly cell-type specific nature of many promoter interactions even with a group of related cell types [51]. Similar to CHi-C, another series of approaches, including Capture-C [83], NG

Capture-C [84] and Tiled-C [85], that focus on capturing chromatin interaction of interest have been developed. Compared to the CHi-C protocols, they enrich the 3C library with biotinylated capture oligonucleotides instead of enrich the biotinylated Hi-C library, allowing the library to retain maximal library complexity, which is important for analysing data from small cell numbers [85].

Like many other high-throughput sequencing approaches, Hi-C continues to be modified to improve the efficiency and resolution of the approach. DNase Hi-C was developed to reduce the bias introduced through the use of restriction enzymes (e.g. MboI recognizes GATC), due to the uneven distribution of restriction sites throughout the genome [77,93]. Instead, DNase Hi-C replaces the restriction enzyme digestion of cross-linked DNA with the endonuclease DNase I that has a much reduced DNA sequence specificity to reduce bias in identifying Hi-C interactions. Commercial Hi-C library preparation kit such as Omni-C kit from Dovetail Genomics [94] exploits the use of DNase and is designed specifically to overcome limitations of only capturing Hi-C interactions near restriction sites. Similar to DNase Hi-C, Micro-C uses micrococcal nuclease (MNase) digestion, enabling the generation of high resolution contact maps at 200 bp to ~4 kb scale in budding yeast [78] and sub-kilobase resolution contact maps in mammalian cells [41,95]. What’s more, BL-Hi-C uses *HaeIII*, which has higher cutting frequency in the human genome compared to other 4-base cutter like *MboI*, to conduct digestion and a two-step ligation optimization to reduce the chance of ligating event of random DNAs, increasing the capture efficiency with active regions in the genome and reducing the probability of random ligation events [79]. In addition to increasing the capture efficiency, optimised protocols are now much more cost effective. For example, DLO Hi-C [80] avoids biotin labeling and pull down steps, and tagHi-C [81] uses Tn5-transposase tagmentation, similar to ATAC-seq, to capture the chromatin structure with hundreds of cells.

The integration of Hi-C with other genomic applications, such as chromatin immunoprecipitation (ChIP), formaldehyde-assisted isolation of regulatory elements (FAIRE) or bisulfite treatment has also occurred. The ChIP-integrated approaches, including HiChIP and PLAC-seq, combining the in situ Hi-C with ChIP, generating a Hi-C library enriched for interactions associated with specific bound proteins [86,87], increasing the resolution of the library while reducing the sequencing cost. Combining the phenol-chloroform extraction step from FAIRE-seq [96] with *in situ* Hi-C, OCEAN-C was developed to prioritise the chromatin interactions on open chromatin [88]. Similarly, integrating with an assay called column purified chromatin (CoP), which is enriched for accessible chromatin regions such as active promoters, enhancers and insulators, HiCoP was recently developed to identify chromatin contacts in regulatory regions [89]. Methyl-HiC has been developed to jointly profile the DNA methylation and 3D genome structure [90]. Recent studies have also revealed that DNA methylation is able to impact 3D genome structure via polycomb complexes, which play an important part in respressing key developmental genes [27,97–100].

The optimizations introduced by protocols such as Micro-C largely improve the crosslinked DNA capture specificity, allowing higher resolution data to be generated with less sequencing cost. Based on these optimizations, Hi-C 2.0 and Hi-C 3.0 have been developed as the updated versions of Hi-C protocol in recent years [91,92]. In Hi-C 3.0, the protocol uses a combination of two restriction enzymes, *DdeI* and *DpnII*, and MNase to generate short fragments, which can improve the identification of genome compartmentalization. Additionally, Hi-C 3.0 also uses DSG as cross-linker in addition to formaldehyde to generate cross-linked DNA, improving the enrichment level of regulatory elements such as promoters and enhancers in the identified chromatin loops [101].

As the development of Hi-C approaches continue, it is essential that computational methods are standardized in order to provide consistent results that are comparable across species or cell-types. In the next section, we review the current data processing methods that are used in standard Hi-C sequencing approaches.

Prioritisation of chromatin interactions

Methodologies to extract meaningful, potentially functional information from the massive number of interactions identified through Hi-C data can be categorized into three groups: structural-based methods, detection of significant interactions and data integration (Figure 2). The first approach is to define structures such as A/B compartments and TADs, based on the 2D interaction patterns across the genome. The second approach is to investigate only a subset of Hi-C interactions that are identified from a statistical test based on a trained model. Finally, taking advantage of the publicly available databases or the generation of epigenomics data in parallel with Hi-C data, the third approach is to prioritise interactions that are more likely to be biologically relevant through the investigation of genomic and epigenomic information. These approaches are not mutually exclusive and in many cases can be combined to address specific questions in genome organisation and gene regulation.

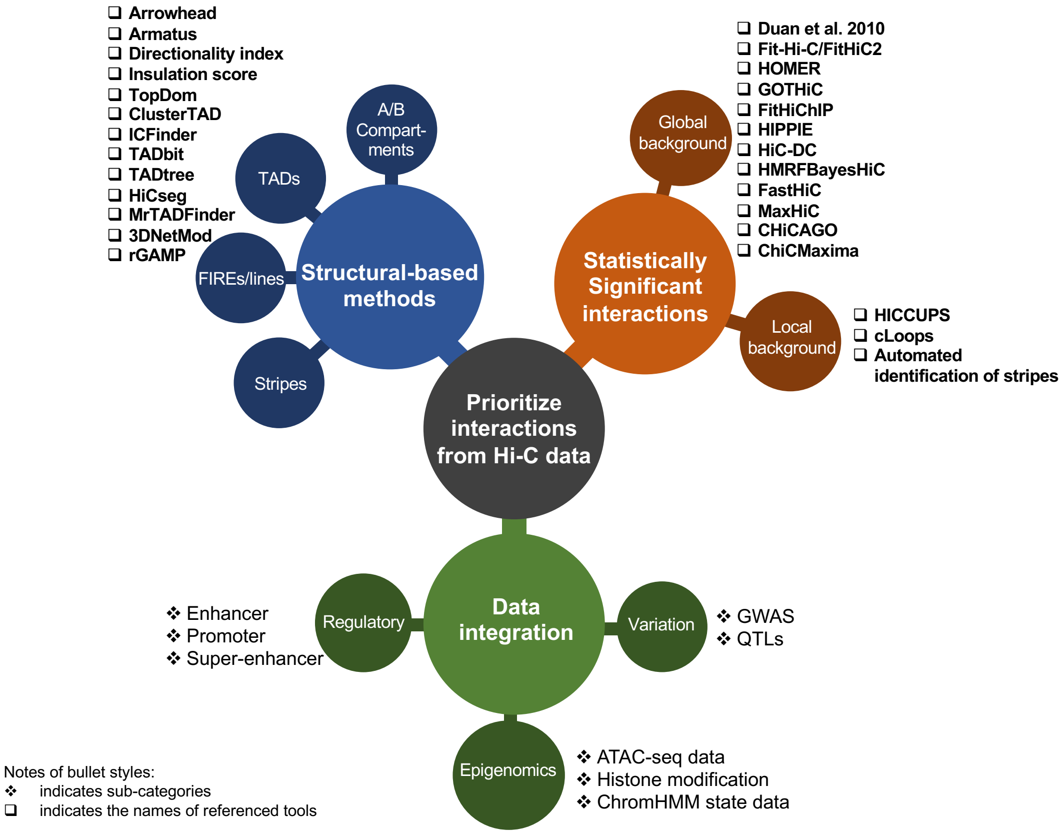


Figure 2: Approaches to prioritise interactions from Hi-C datasets. In this review, we categorised the approaches to identify potentially functional interactions into three ways, including significant interactions identification, structures summarisation and data integration. Referenced tools and sub-categorical analyses are marked on the figure with boxes and stars respectively.

Structural-based identification methods

Methods that identify structural aspects of chromatin interactions (i.e. A/B compartments and TADs) are employed as an avenue to reduce the dimensionality of the 3D interaction patterns across the genome by clustering or summarising regions with similar patterns across the genome. The A/B compartments are commonly predicted with normalised Hi-C matrices generated using vanilla coverage (VC) [2], Knight and Ruiz's method (KR) [102] or iterative correction and eigenvector decomposition (ICE) [103]. Normalised data is then used to calculate Pearson's correlation and through principal component analysis (PCA), the eigenvectors of the first (or second) principal component (PC) are usually used to assign bins to A or B compartments. Current analysis toolkits, such as Juicer [104] and FAN-C [105], have optimised correlation matrix functions to identify A/B compartments from Hi-C matrices without significant taxes on memory and computational resources.

As detailed above, TADs are defined as structures with interactions that occur within TADs rather than across TADs [30]. As such, they are often identified by finding domains where contacts are enriched within the same TAD as compared to neighboring TADs [30,106]. Currently, there are over 20 commonly used TADs callers that have been developed using various methodologies. For instance, arrowhead [3], armatus [107], directionality index [30], insulation score [108] and TopDom [109] use their own linear scoring system, clusterTAD [110] and ICFinder [111] are based on clustering, TADbit [112], TADtree [113] and HiCseg [114] use statistical models; and MrTADFinder [115] and 3DNetMod [116] rely on network-modelling approaches [37,117]. Although comparisons reveal low reproducibility among tools, especially in the number and mean size of identified TADs, recent reviews [37,117] have suggested a preference for TAD callers that allow for the detection of nested TADs or overlapped TADs, such as rGMAP [118], armatus, arrowhead and TADtree.

While theoretically similar to TAD calling, frequently interacting regions (FIREs) are also commonly used to describe structural interaction characteristics. Defined as genomic regions with significant interaction profile, FIREs exhibit strong connectivity with multiple regions in the chromosome neighbourhood [73]. FIREs can be easily visualised on the Hi-C interaction map, with interacting signals appearing from both sides of the FIREs, forming a characteristic “V” shape (Figure 1A). Unlike TADs and compartments, which exhibit a certain level of conservation across cell types (about 50~60% and 40%, respectively) [3,30,73,119], FIREs appear to be cell type- and tissue-specific and are often located near key cell phenotype-defining genes. However, similar to TADs, FIREs formation seems to be dependent on the Cohesin complex, as its depletion results in decreasing interactions at FIREs [73]. They are also enriched for super-enhancers, suggesting FIREs play an important role in the dynamic gene regulation network [120,121]. Similar to FIREs, “V” shape structural feature that is referred to as “line” structure was observed at the edge of the TADs during the exploration or loop extrusion model using simulated Hi-C data [14].

Methods for identification of significant chromatin interactions

In order to prioritize potentially meaningful chromatin interactions, statistical significance is assigned to Hi-C interactions by comparing them to a background model and assessing the probability of observing the experimental set of counts if the background model were the underlying method of generating observed counts. The interaction frequency generally decays with increasing linear distance, and by applying this background model meaningful interactions can be identified through a higher than normal frequency. Here we summarize the current methodologies of significant interactions identification and categorise them into two groups; global background model methods, which define a background signal model by considering the read count of any pair of interactions, and local background model methods, which account for interactions in the neighbouring areas to identify peak interactions with statistical significance.

Table 2: methods for identification of statistically significant interactions for Hi-C data.

Method name	Type	Base model	Specific features	Reference

Duan et al. 2010	Global background	Binomial	Specifically designed for yeast genome	[122]
Fit-Hi-C/FitHiC2	Global background	Binomial	Spline fitting procedure, compatible with different formats	[123,124]
HOMER	Global background	Binomial	Highly compatible with the HOMER Hi-C analysis pipeline	[125]
GOTHiC	Global background	Binomial	Use relative coverage to estimate biases	[126]
FitHiChIP	Global background	Binomial	Specifically designed for HiChIP data	[127]
HIPPIE	Global background	Negative binomial	Account for fragment length and distance biases	[72,128]
HiC-DC	Global background	Negative binomial	Use zero-inflated model	[129]
HMRFBayesHiC	Global background	Negative binomial	Use hidden Markov random field model	[130]
FastHiC	Global background	Negative binomial	An updated version of HMRFBayesHi, with improved computing speed	[131]
MaxHiC	Global background	Negative binomial	Use ADAM algorithm, identify interactions with enrichment for regulatory elements	[132]
CHiCAGO	Global background	Negative binomial	Specifically designed for CHi-C data	[133]
ChiCMaxima	Global background	Local maxima	Specifically designed for CHi-C data, more stringent and robust when comparing biological replicates	[134]
HICCUP	Local background	Local enrichment	Robust for finding chromatin loops	[3]
cLoops	Local background	DBSCAN	Loop detection with less computational resource	[135]
Automated identification of stripes	Local background	Local enrichment	Specifically designed to identify architectural stripes	[136]

Global background-based methods

The Initial study which assigns statistical significance to Hi-C interactions is done in the yeast genome. The chromatin interactions in the yeast genome was first separated into intra-chromosomal interactions (within the same chromosome) and inter-chromosomal interactions (across two chromosomes), followed by a binomial distribution to assign confidence estimates for inter-chromosomal interactions [122]. A binning method is then used to account for the characteristic pattern of intra-chromosomal interactions, with the observed interacting probability decaying as the genomic distance increases linearly. This is then used to compute interacting probabilities for each bin separately and assigning statistical significance using the same binomial distribution as used for inter-chromosomal interactions [122]. Based on the same binomial distribution concept, Fit-Hi-C uses spline fitting procedure instead of binning, reducing the bias of artifactual stair-step pattern, allowing detection of statistically significant interactions in the mammalian genome [123]. Additionally, Fit-Hi-C also incorporates an extra refinement step using a conservative model with stringent parameters to remove outlier interactions, which can be applied iteratively, to achieve a more accurate empirical null model. However, Fit-Hi-C was initially limited by only allowing bin sizes larger than 5 kb to compute significance due to the heavy memory usage when dealing with higher-resolution data. However this has been improved with recent updates [124], and is now able to handle data with high resolution (bin sizes from 1 to 5 kb). Another important new feature is that it is now accepting multiple input formats so that it is compatible with different Hi-C analysis pipelines. Another similar tool is included in the Homer toolkit [125], which accounts for biases such as sequencing depths, linear distance between regions, GC bias and chromatin compaction to establish a background model to estimate the expected interaction count between any two regions, followed by the use of a cumulative binomial distribution to assign significance to interactions. GOTHiC [126] also uses relative coverage of two interacting regions to estimate both known and unknown biases, followed by a cumulative binomial distribution to build the background model to identify significant interactions

The Negative Binomial distribution is commonly utilised in the analysis of count-based data, including popular RNA-seq analysis tools such as edgeR [137] and DEseq2 [138], and has been implemented in a number of Hi-C programs such as HIPPIE [72,128]. This method uses a negative binomial model to estimate the statistical significance of the interactions in one fragment region (< 2 Mb) while accounting for restriction fragment length bias and interacting probability distance bias simultaneously. However, negative binomial models can be confounded by many bins with zero counts [129] and a number of programs have developed approaches to account for “zero-inflated” observations. HiC-DC, for example, uses a hurdle negative binomial regression model to identify significant interactions [129], modelling the probability of non-zero counts and the rate of observed counts as separate components of the model.

While physical interactions between loci found in close linear proximity are likely to be more prevalent in Hi-C datasets, a known bias in Hi-C libraries is the correlation between two nearby restriction fragments brought about by ligation events. Ligation events can be the result of bias or random collision events between restriction fragments during library preparation, so with high coverage sequencing, false signals can impact the identification of significant interactions [72]. To tackle this problem, HMRFBayesHiC uses a negative binomial distribution to model observed interactions [72], followed by a hidden Markov random field model to account for the correlation between restriction fragments, and to model interaction probabilities [130]. This implementation required significant resources to run, leading to the development of FastHiC [131], which enables higher accuracy of interaction identification and faster performance. Recently, another tool called MaxHiC also based on negative binomial distribution was developed [132]. Compared to other tools, all parameters of the background model in MaxHiC are established by using the ADAM algorithm [139] to maximize the logarithm of likelihood of the observed Hi-C interactions. Significant interactions identified by MaxHiC were shown to outperform tools such as Fit-Hi-C/FitHiC2 and GOTHiC in identifying significant interactions enriched between known regulatory regions [132].

Compared to traditional Hi-C protocols, Capture Hi-C (CHi-C) requires different analytic methods due to the extra bias driven by the enrichment step in the protocol. Capture libraries can be regarded as a subset of the original Hi-C library, meaning the interaction matrix of CHi-C is asymmetric, and interestingly not accounted for in traditional normalisation methods [82,133]. Because of this, many analysis approaches are specifically designed for CHi-C data analysis. CHiCAGO (Capture Hi-C Analysis of Genomic Organisation) was developed to account for biases from the CHi-C protocol and identify significant interactions [133], using a negative binomial distribution to model the background local profile and an additional Poisson random variable to model technical artefacts [133]. CHiCAGO uses the implicit normalization method ICE [103] and multiple testing stages based on p-value weighting [140] to carefully identify significant interactions from each CHi-C dataset [133]. Another CHi-C-specific tool called ChiCMaxima was developed to identify significant interactions by defining them as local maxima after using loess smoothing on bait-specific interactions [134]. Compared to CHiCAGO, ChiCMaxima's approach is more stringent and exhibits a more robust performance when comparing biological replicates [134]. As well as being applicable to conventional HiC data, MaxHiC is also able to identify significant interactions in CHi-C data [132] and offers robust performance to identify regulatory areas compared to CHi-C-specific tools including CHiCAGO [132].

Like the other capture approaches, HiChIP cannot use traditional (Hi-C-specific) interaction callers (e.g. Fit-Hi-C or GOTHiC) due to the inherent biases associated with an enrichment with specific immunoprecipitation targets [86]. Hichipper was developed to firstly identify ChIP peaks while accounting for the read density bias in restriction fragments, enabling a more accurate identification of interactions from HiChIP dataset [141]. While hichipper does not implement any function to identify significant interactions, FitHiChIP was developed to account for non-uniform coverage bias and distance bias in restriction fragments using a regression model, together with 1D peak information in a spline fitting procedure to accurately identify significant interactions from HiChIP data [127].

Local background-based methods

Chromatin looping structures can be regarded as the basic unit of 3D genomic architecture and play an important role in the regulatory process, by bringing distal promoter and enhancer elements together or excluding enhancers from the looping domain [15–17]. Chromatin loops from Hi-C data were first defined by searching for the strongest “pixel” on a normalised Hi-C contact map (Figure 1A). Different from the global background models used by methods like Fit-Hi-C and MaxHiC, using a local background model to compare all pixels in a neighbouring area is able to detect pixels with the strongest signals as the anchor points of chromatin loops [3]. A searching algorithm named Hi-C Computational Unbiased Peak Search (HICCUPS) was therefore developed to rigorously search for these pixels based on the local enrichment in the pixel neighborhood, followed by hypothesis testing with Poisson statistics, enabling the identification of chromatin loops from Hi-C data [3]. Somewhat similar to TADs, published information on chromatin loops demonstrates structural conservation between a number of human cell lines (~55-75% similarity), and between human and mouse (about 50% similarity), suggesting conserved loops may serve as a basic functional unit for the genome [3]. However, loop detection using HICCUPS requires high resolution data with extremely high sequencing depth. For example, almost 5 billion unique interactions were required by HICCUPS to identify 10,000 unique loops in the GM12878 cell line [3]. This limitation can potentially be addressed by the current development of deep learning approaches, such as DeepHiC [142] using generative adversarial networks, as well as HiCPlus [143] and HiCNN [144] which use deep convolutional neural networks. Such methods can be used to increase the resolution of Hi-C data to achieve necessary resolution so that chromatin loops can be identified, or to improve loop detection accuracy [142,143].

Hardware requirements to identify loops in high-resolution data is also extremely restrictive with HICCUPS requiring specific architectures (i.e. NVIDIA GPUs) to identify looping patterns. However this has been addressed recently with the HICCUPS algorithm being reimplemented in the cooltools package (https://github.com/mirnylab/cooltools), allowing HICCUPS to be run on a regular server or compute cluster [95]. Alternatively, an approach called cLoops was implemented which identifies peak interactions from chromatin contact map [135]. cLoops initiates loop detection by finding candidate loops via an unsupervised clustering algorithm, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [145], which enables computing statistical significance of interactions with less amount of input and reduced computational resources. Candidate loops are then compared with a permuted background model, based on the interaction decay over linear distance, to estimate statistical significance.

Further investigation in high-resolution Hi-C data (<= 10 kb), another local background model method was developed to identify architectural stripe structures rather than loops [136]. The stripe structure is similar to FIRE, where a genomic region contacts other regions of the entire domain with high interacting frequency [136]. Its identification algorithm *Automated identification of stripes* computes the pixel-specific enrichment relative to its local neighbourhood, then performs Poisson statistics to test if the signal is statistically significant [136]. It was further shown that stripe anchors highly correspond to loop anchors, and stripes appear to be relevant with enhancer activity [136,146].

Potentially functional interaction identification via data integration

While variation in gene-coding regions can lead to significant alterations in one gene or abnormalities across a region in the genome, causing mendelian diseases such as chronic granulomatous disease [147], cystic fibrosis [148] and Fanconi's anaemia [149]. The fundamental motivation for identifying interacting regions across a genome is to establish how non-coding regions of the genome impact gene expression [1,150,151]. However potentially functionally-relevant interactions, whether this be chromatin interactions between gene promoters and enhancers or transcription factor binding mechanisms, are often established in a cell-type specific manner [71,82]. By integrating Hi-C interactions with local or publicly available genomic, transcriptomic and epigenomic datasets, such as regulatory elements, gene expression, genetic variation and quantitative trait loci (QTL) information, potentially functional interactions can be prioritised.

Potentially functional Hi-C interactions can be identified by integration with transcriptomics and enhancer data. Promoter-enhancer interactions (PEI), promoter-promoter interactions (PPI) or enhancer-enhancer interactions (EEI), where distal promoter or enhancer are brought into close proximity by chromatin contacts to form complex contact, are three widely accepted potentially functional Hi-C interaction types to be studied [51,69,152–158]. These interaction categories are often identified by finding overlaps of promoter or enhancer signals separately at each anchor of a Hi-C interaction [51,156,159]. However, when identifying PEI or PPI from Hi-C data for a specific cell type, the gene expression profile of such cell type should be considered to determine which promoters are active given that promoter interactions are shown highly cell-type specific [51].

Similar to promoters of expressed genes, active enhancers of a specific cell type are necessary to identify potentially functional PEI or EEI for a specific cell type. Expressed enhancers (eRNAs) or experimentally verified enhancers of different human cell types and tissues are available in publicly available projects and databases such as FANTOM5 project [160], the NIH Roadmap Epigenomics project [161], the EU Blueprint project [162], ENCODE [163,164] and ENdb [165]. Additionally, previous studies also used cell type-specific histone markers ChIP-seq data, such as H3K27ac and H3K4me1, or integrated chromHMM chromatin state information predicted from a variety of epigenomic sequencing information [166,167] to indicate the activity of an enhancer in a specific cell type

[51,156,159,168,169]. In addition to using Hi-C data, there are numerous methods that have been developed to predict potentially functional interactions based on histone marker signals [170], gene expression and methylation data [171], ATAC-seq data [172], DNase-seq data [173] or even DNA sequence alone [174]. These types of methods have been comprehensively reviewed in a recent review study [175].

Besides promoters and enhancers, Super-enhancers (SEs) are another major regulatory element that is crucial to the identification of potentially functional interactions. SEs are defined as a clustered region of enhancers exhibiting significantly higher levels of active enhancer marks and an enrichment with transcription factor binding sites (TFBS) [176]. These regions act as “regulatory hubs”, which are higher-order complexes consisting of interactions between multiple enhancers and promoters at individual alleles [153,177,178]. The formation of these regulatory hubs are proposed to be the consequence of the high level of TF and co-factor localisation to the SE interacting to form a biomolecular condensate by a phase separation model [179–184]. Identified Hi-C interactions with linkages to SE have been shown to be potentially functional by mediating multiple gene expression regulations three-dimensionally, or being essential for cell identity and development [50,185–190]. SE can be identified from H3K27ac ChIP-seq using the ROSE algorithm [187], and currently SE information can be easily accessible from databases such as AnimalTFDB [191], PlantTFDB [192], GTRD [193], SEdb [194], dbSUPER [195] and SEA [196,197], allowing cell-type regulatory hubs to identified and linked to phenotypic traits and/or disease.

In genome-wide association studies (GWAS), almost 90% of the identified genetic single nucleotide polymorphisms (SNPs) associated with phenotypic traits are located in non-coding regions such as gene desert, which are areas lacking protein-coding genes, hence making the interpretation of the functions of such variants much more challenging than the ones located within or nearby protein-coding genes [198–200]. Hi-C data have been proved to be useful in many studies for addressing this issue by forming linkages between diseases-associated variants and genes using long-range chromatin interactions. For examples, interactions between gene promoters and variation-located long coding RNAs (lncRNA), where GWAS SNPs can impact the expression of the target genes by affecting the binding of TF binding to the lncRNA [201]; direct interactions between SNPs and multiple genes, exhibiting co-regulation function of the SNPs [202]; interaction networks based on a SNP, bringing gene promoter, TF binding site and active enhancer region together by chromatin interactions to affect gene expression [203]. Variants may also impact gene-coding regions over large distances meaning that target genes of the variations are not necessarily their closest proximal gene [71,204]. Currently, databases such as GWAS catalog [205], ImmunoBase [206], GWAS Central [207], GWAS ALTAS [208] and GWASdb [209] contain information of the level of genetic association of each variant to specific diseases, which are invaluable data to be integrated in a high-dimensional interaction dataset.

Tissue-specific quantitative trait loci (QTLs) are identified as the possession of variants that can significantly impact the level of quantitative trait [210], such as expression QTLs (eQTLs) that affect the expression level of the target genes [211], histone QTLs (hQTLs) that affect histone modifications [212,213], methylation QTLs (meQTLs) that impact DNA methylations [214,215] and ATAC-QTL that affect the accessibility of the corresponding areas [216]. In recent QTL studies, QTLs are found to affect their target regions by the long-range chromatin interactions between them observed from Hi-C data. For example, Greenwald *et al.* has recently used pancreatic islet-specific data to investigate the risk gene loci of Type 2 Diabetes (T2D) [217]. In their work they combined gene and enhancers interaction maps generated from Hi-C data, together with variant and gene expression linkage data, provided by tissue-specific eQTL analysis, to establish an enhancer network for T2D risk loci. In support of genetic variation at enhancers influencing transcriptional regulation, Yu *et al.* used HiC data to demonstrate that eQTLs tend to be in close spatial proximity with their target genes [218]. Additionally, a recent multi-tissues integration analysis between eQTLs and Hi-C interactions revealed the close proximity between eQTLs and their target genes, indicating that eQTLs regulate the expression of their target genes through chromatin contacts [218]. Therefore, with publicly available QTL databases such as the GTEx project [211], seeQTL [219], Haploreg [220], Blood eQTL browser [221], Pancan-meQTL [222] and QTLbase [223], the linkages between such QTLs and their target genes or regions can be used to infer potentially functional Hi-C interactions.

Future prospects

The investigation of 3D chromosome structure can provide novel insights into the complex regulatory network in the genome. The development of Hi-C and its derived protocols have facilitated the studies of the 3D genome structure, generating numerous high quality datasets. However, due to the complexity of the Hi-C library preparation and analysis, the biologically meaningful, small-scale interactions may still lack sufficient signals, hindering the detection and interpretation of 3D interactions. The approaches that we presented in this review all aim to reduce the complexity of 3D interaction data, narrowing down information based on structure, statistical inference and additional lines of experimental evidence (i.e. cell-type specific epigenomic data).

Incremental development of Hi-C calling applications (chromatin loops, TADs etc) has continued with a focus on correcting biases introduced by library preparation and sequencing. As more and more sequencing data is deposited on open-access data repositories such as NCBI Short Read Archive (SRA) [224] and European Nucleotide Archive (ENA) [225], has allowed the development of novel Machine Learning models trained on known interactions to identify novel patterns when applying these models to new datasets. Incorporation of publicly-available cell type/tissue-specific epigenomics data into these machine learning models of chromatin interactions will allow for more accurate predictions on the molecular mechanisms by which diseases-associated genetic acts. In the future, such models of 3D interactions can potentially be used as markers for disease screening and used for personalised medicine development.

Although the development in protocol efficiency, parallel algorithmic improvements are likely to improve current approaches for identifying 3D interactions. Additional imaging technologies such as real-time signal Fluorescence in situ hybridization and advanced imaging approaches such as STORM imaging have been used to visualise the nuclear organization in living cells and leading to the identification of clusters of clutch domains that are thought to correspond to TAD [7,226]. Lastly the ability to engineer specific mutations in DNA through genome editing technology such as the CRISPR-Cas9 system [227,228], means that future experiments using Hi-C and 3D imaging in-parallel with genetically modification of genomes will vastly improve our understanding of how variation may impact genomic structure, and the regulations of gene expression.

Conclusion

In this review, we first introduced the three-dimensional chromosome architecture in different scales, followed by presenting the chromosome conformation capture assays, with a focus on Hi-C and its variations, which are the state of the art methods for investigating the 3D genome structure. Lastly, we comprehensively reviewed methodologies that are developed to reduce the complexity of 3D physical interactions identified from Hi-C datasets to detect potentially functional interactions. We also categorised the methods into three types, including structural-based detection methods, significant chromatin interactions identification methods and data integration methods. Taken together, by utilizing these methods carefully, we are able to detect physical interactions with biological meaning and impact from complicated Hi-C dataset, which may serve a purpose in diagnosis and precision medicine.

List of Abbreviations

CTCF: CCCTC-binding protein

TAD: Topologically-associated domain

FIRE: Frequently interacting region

NADs: Nucleolus-associated domains

3D-FISH: 3D Fluorescence *in situ* Hybridisation

3C: Chromosome conformation capture

4C-seq: Circular chromosome conformation capture sequencing

5C: Chromosome conformation capture carbon copy

ChIA-PET: Chromatin interaction analysis by paired-end tag sequencing

Hi-C: Higher-resolution chromosome conformation capture sequencing

CHi-C: Capture Hi-C

PEI: Promoter-enhancer interactions

PPI: Promoter-promoter interactions

EEI: Enhancer-enhancer interactions

SE: Super-enhancers

TFBS: Transcription factor binding sites

lncRNA: Long coding RNAs

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by a 2017 National Health and Medical Research Council (NHMRC) project grant (*1120543*).

Authors' contributions

JB and NL designed this review with input from all other authors. NL wrote the manuscript in conjunction with WYL, HA and JB. All authors edited and approved the manuscript.

Acknowledgements

We are thankful for the access to all publicly available tools and software packages.

References

1. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002;295:1306–11.

2. 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;326:289–93.

3. Rao SSP, 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;159:1665–80.

4. Taberlay PC, Achinger-Kawecka J, Lun ATL, Buske FA, Sabir K, Gould CM, et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. Genome Res. 2016;26:719–31.

5. Anania C, Lupiáñez DG. Order and disorder: abnormal 3D chromatin organization in human disease. Brief Funct Genomics. 2020;19:128–38.

6. Liu N, Sadlon T, Wong YY, Pederson SM, Breen J. 3DFAACTS-SNP: Using regulatory T cell-specific epigenomics data to uncover candidate mechanisms of Type-1 Diabetes (T1D) risk. bioRxiv [Internet]. biorxiv.org; 2020; Available from: https://www.biorxiv.org/content/10.1101/2020.09.04.279554v1.abstract

7. Lakadamyali M, Cosma MP. Visualizing the genome in high resolution challenges our textbook understanding. Nat Methods. Nature Publishing Group; 2020;17:371–9.

8. Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. Nat Rev Genet. 2018;19:789–800.

9. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet. Springer Science and Business Media LLC; 2016;17:661–78.

10. Maass PG, Barutcu AR, Rinn JL. Interchromosomal interactions: A genomic love story of kissing chromosomes. J Cell Biol. 2019;218:27–38.

11. Davidson IF, Bauer B, Goetz D, Tang W, Wutz G, Peters J-M. DNA loop extrusion by human cohesin. Science. 2019;366:1338–45.

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

13. Sanborn AL, Rao SSP, Huang SC, Durand NC, Huntley MH, Jewett AI, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. PNAS, 112. E6456--E6465. 2015;

14. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. 2016;15:2038–49.

15. Kadauke S, Blobel GA. Chromatin loops in gene regulation. Biochim Biophys Acta. 2009;1789:17–25.

16. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, et al. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell. Elsevier; 2015;162:900–10.

17. Krijger PHL, de Laat W. Regulation of disease-associated gene expression in the 3D

genome. Nat Rev Mol Cell Biol. 2016;17:771–82.

18. Splinter E, Heath H, Kooren J, Palstra R-J, Klous P, Grosveld F, et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev. 2006;20:2349–54.

19. Rubio ED, Reiss DJ, Welcsh PL, Disteche CM, Filippova GN, Baliga NS, et al. CTCF physically links cohesin to chromatin. Proc Natl Acad Sci U S A. 2008;105:8309–14.

20. Zuin J, Dixon JR, van der Reijden MIJA, Ye Z, Kolovos P, Brouwer RWW, et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc Natl Acad Sci U S A. 2014;111:996–1001.

21. Banigan EJ, van den Berg AA, Brandão HB, Marko JF, Mirny LA. Chromosome organization by one-sided and two-sided loop extrusion. Elife [Internet]. 2020;9. Available from: http://dx.doi.org/10.7554/eLife.53558

22. Banigan EJ, Mirny LA. Loop extrusion: theory meets single-molecule experiments. Curr Opin Cell Biol. 2020;64:124–38.

23. Drissen R, Palstra R-J, Gillemans N, Splinter E, Grosveld F, Philipsen S, et al. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. Genes Dev. 2004;18:2485–90.

24. Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, et al. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. Mol Cell. 2005;17:453–62.

25. Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell. 2012;149:1233–44.

26. Apostolou E, Ferrari F, Walsh RM, Bar-Nur O, Stadtfeld M, Cheloufi S, et al. Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. Cell Stem Cell. 2013;12:699–712.

27. Denholtz M, Bonora G, Chronis C, Splinter E, de Laat W, Ernst J, et al. Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. Cell Stem Cell. 2013;13:602–16.

28. Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell. 2017;171:1573–88.e28.

29. Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. Cell Syst. 2018;6:256–8.e1.

30. 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;485:376–80.

31. Tang Z, Luo O, Li X, Zheng M, Zhu J, Szalaj P, et al. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell.

2015;163:1611–27.

32. de Wit E, Vos ESM, Holwerda SJB, Valdes-Quezada C, Verstegen MJAM, Teunissen H, et al. CTCF Binding Polarity Determines Chromatin Looping. Mol Cell. 2015;60:676–84.

33. Beagan JA, Phillips-Cremins JE. On the existence and functionality of topologically associating domains. Nat Genet. 2020;52:8–16.

34. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. Sci Adv. 2019;5:eaaw1668.

35. Wit E de. TADs as the caller calls them. J Mol Biol [Internet]. 2019; Available from: http://dx.doi.org/10.1016/j.jmb.2019.09.026

36. Rocha PP, Raviram R, Bonneau R, Skok JA. Breaking TADs: insights into hierarchical genome organization. Epigenomics. 2015;7:523–6.

37. Zufferey M, Tavernari D, Oricchio E, Ciriello G. Comparison of computational methods for the identification of topologically associating domains. Genome Biol. 2018;19:217.

38. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. 2015;10:1297–309.

39. Rowley MJ, Nichols MH, Lyu X, Ando-Kuri M, Rivera ISM, Hermetz K, et al. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. Mol Cell. 2017;67:837–52.e7.

40. Llères D, Moindrot B, Pathak R, Piras V, Matelot M, Pignard B, et al. CTCF modulates allele-specific sub-TAD organization and imprinted gene activity at the mouse Dlk1-Dio3 and Igf2-H19 domains. Genome Biol. 2019;20:272.

41. Hsieh T-HS, 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;78:539–53.e8.

42. Giorgetti L, Lajoie BR, Carter AC, Attia M, Zhan Y, Xu J, et al. Structural organization of the inactive X chromosome in the mouse. Nature. 2016;535:575–9.

43. Wang Q, Sun Q, Czajkowsky DM, Shao Z. Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. Nat Commun. 2018;9:188.

44. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, et al. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012;488:116–20.

45. Nora EP, Dekker J, Heard E. Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? Bioessays. 2013;35:818–28.

46. Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, et al. Functional and topological characteristics of mammalian regulatory domains. Genome Res.

2014;24:390–400.

47. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161:1012–25.

48. Smith EM, Lajoie BR, Jain G, Dekker J. Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. Am J Hum Genet. 2016;98:185–201.

49. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science. 2016;351:1454–8.

50. Huang J, Li K, Cai W, Liu X, Zhang Y, Orkin SH, et al. Dissecting super-enhancer hierarchy based on chromatin interactions. Nat Commun. 2018;9:943.

51. 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;167:1369–84.e19.

52. Montavon T, Soshnikova N, Mascrez B, Joye E, Thevenet L, Splinter E, et al. A regulatory archipelago controls Hox genes transcription in digits. Cell. 2011;147:1132–45.

53. Laat W de, de Laat W, Klous P, Kooren J, Noordermeer D, Palstra R, et al. Chapter 5 Three-Dimensional Organization of Gene Expression in Erythroid Cells [Internet]. Red Cell Development. 2008. p. 117–39. Available from: http://dx.doi.org/10.1016/s0070-2153(07)00005-1

54. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell. 2002;10:1453–65.

55. Palstra R-J, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, de Laat W. The β-globin nuclear compartment in development and erythroid differentiation [Internet]. Nature Genetics. 2003. p. 190–4. Available from: http://dx.doi.org/10.1038/ng1244

56. Noordermeer D, de Laat W. Joining the loops: beta-globin gene regulation. IUBMB Life. Wiley; 2008;60:824–33.

57. Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature. 2011;472:120–4.

58. Kim YJ, Cecchini KR, Kim TH. Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. Proc Natl Acad Sci U S A. 2011;108:7391–6.

59. Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D. The dynamic architecture of Hox gene clusters. Science. 2011;334:222–5.

60. Noordermeer D, Leleu M, Schorderet P, Joye E, Chabaud F, Duboule D. Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci.

Elife. 2014;3:e02557.

61. Di Giammartino DC, Polyzos A, Apostolou E. Transcription factors: building hubs in the 3D space. Cell Cycle. 2020;19:2395–410.

62. Rickels R, Shilatifard A. Enhancer Logic and Mechanics in Development and Disease. Trends Cell Biol. 2018;28:608–30.

63. Smith E, Shilatifard A. Enhancer biology and enhanceropathies. Nat Struct Mol Biol. 2014;21:210–9.

64. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet. 2006;38:1348–54.

65. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006;16:1299–309.

66. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009;462:58–64.

67. Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet. 2006;38:1341–7.

68. Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M, et al. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic Acids Res. 2010;38:8164–77.

69. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489:109–13.

70. Rodley CDM, Bertels F, Jones B, O’Sullivan JM. Global identification of yeast chromosome interactions using Genome conformation capture. Fungal Genet Biol. 2009;46:879–86.

71. Martin P, McGovern A, Orozco G, Duffus K, Yarwood A, Schoenfelder S, et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. Nat Commun. nature.com; 2015;6:10069.

72. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013;503:290–4.

73. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep. 2016;17:2042–59.

74. Barutcu AR, Hong D, Lajoie BR, McCord RP, van Wijnen AJ, Lian JB, et al. RUNX1 contributes to higher-order chromatin organization and gene regulation in breast cancer cells. Biochim Biophys Acta. 2016;1859:1389–97.

75. van de Werken HJG, Landan G, Holwerda SJB, Hoichman M, Klous P, Chachik R, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods. 2012;9:969–72.

76. Nagano T, Lubling Y, Yaffe E, Wingett SW, Dean W, Tanay A, et al. Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. Nat Protoc. 2015;10:1986–2003.

77. Ma W, Ay F, Lee C, Gulsoy G, Deng X, Cook S, et al. Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. Nat Methods. 2015;12:71–8.

78. Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell. 2015;162:108–19.

79. Liang Z, Li G, Wang Z, Djekidel MN, Li Y, Qian M-P, et al. BL-Hi-C is an efficient and sensitive approach for capturing structural and regulatory chromatin interactions. Nat Commun. 2017;8:1622.

80. Lin D, Hong P, Zhang S, Xu W, Jamal M, Yan K, et al. Digestion-ligation-only Hi-C is an efficient and cost-effective method for chromosome conformation capture. Nat Genet. 2018;50:754–63.

81. Zhang C, Xu Z, Yang S, Sun G, Jia L, Zheng Z, et al. tagHi-C Reveals 3D Chromatin Architecture Dynamics during Mouse Hematopoiesis. Cell Rep. 2020;32:108206.

82. 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;47:598–606.

83. Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet. 2014;46:205–12.

84. Davies JOJ, Telenius JM, McGowan SJ, Roberts NA, Taylor S, Higgs DR, et al. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. Nat Methods. 2016;13:74–80.

85. Oudelaar AM, Beagrie RA, Gosden M, de Ornellas S, Georgiades E, Kerry J, et al. Dynamics of the 4D genome during in vivo lineage specification and differentiation. Nat Commun. 2020;11:2722.

86. 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;13:919–22.

87. Fang R, Yu M, Li G, Chee S, Liu T, Schmitt AD, et al. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. Cell Res. 2016;26:1345–8.

88. Li T, Jia L, Cao Y, Chen Q, Li C. OCEAN-C: mapping hubs of open chromatin

interactions across the genome reveals gene regulatory networks. Genome Biol. 2018;19:54.

89. Zhang Y, Li Z, Bian S, Zhao H, Feng D, Chen Y, et al. HiCoP, a simple and robust method for detecting interactions of regulatory regions. Epigenetics Chromatin. 2020;13:27.

90. Li G, Liu Y, Zhang Y, Kubo N, Yu M, Fang R, et al. Joint profiling of DNA methylation and chromatin architecture in single cells. Nat Methods. 2019;16:991–3.

91. Belaghzal H, Dekker J, Gibcus JH. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. Methods. 2017;123:56–65.

92. Oksuz BA, Yang L, Abraham S, Venev SV. Systematic evaluation of chromosome conformation capture assays. bioRxiv [Internet]. biorxiv.org; 2020; Available from: https://www.biorxiv.org/content/10.1101/2020.12.26.424448v1.abstract

93. Ramani V, Cusanovich DA, Hause RJ, Ma W, Qiu R, Deng X, et al. Mapping 3D genome architecture through in situ DNase Hi-C. Nat Protoc. 2016;11:2104–21.

94. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 2016;26:342–50.

95. Krietenstein N, Abraham S, Venev SV, Abdennur N, Gibcus J, Hsieh T-HS, et al. Ultrastructural Details of Mammalian Chromosome Architecture. Mol Cell. 2020;78:554–65.e7.

96. Simon JM, Giresi PG, Davis IJ, Lieb JD. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nat Protoc. 2012;7:256–67.

97. Schoenfelder S, Sugar R, Dimond A, Javierre B-M, Armstrong H, Mifsud B, et al. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nat Genet. 2015;47:1179–86.

98. Vieux-Rochas M, Fabre PJ, Leleu M, Duboule D, Noordermeer D. Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. Proc Natl Acad Sci U S A. 2015;112:4672–7.

99. Joshi O, Wang S-Y, Kuznetsova T, Atlasi Y, Peng T, Fabre PJ, et al. Dynamic Reorganization of Extremely Long-Range Promoter-Promoter Interactions between Two States of Pluripotency. Cell Stem Cell. 2015;17:748–57.

100. McLaughlin K, Flyamer IM, Thomson JP, Mjoseng HK, Shukla R, Williamson I, et al. DNA Methylation Directs Polycomb-Dependent 3D Genome Re-organization in Naive Pluripotency. Cell Rep. 2019;29:1974–85.e6.

101. Oksuz BA, Yang L, Abraham S, Venev SV. Systematic evaluation of chromosome conformation capture assays. bioRxiv [Internet]. biorxiv.org; 2020; Available from: https://www.biorxiv.org/content/10.1101/2020.12.26.424448v1.abstract

102. Knight PA, Ruiz D. A fast algorithm for matrix balancing. IMA J Numer Anal. Narnia; 2013;33:1029–47.

103. 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;9:999–1003.

104. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 2016;3:95–8.

105. Kruse K, Hug CB, Vaquerizas JM. FAN-C: A Feature-rich Framework for the Analysis and Visualisation of C data. bioRxiv [Internet]. biorxiv.org; 2020; Available from: https://www.biorxiv.org/content/10.1101/2020.02.03.932517v1.abstract

106. Chang L-H, Ghosh S, Noordermeer D. TADs and Their Borders: Free Movement or Building a Wall? J Mol Biol. 2020;432:643–52.

107. Filippova D, Patro R, Duggal G, Kingsford C. Identification of alternative topological domains in chromatin. Algorithms Mol Biol. 2014;9:14.

108. Crane E, Bian Q, Rachel M, Lajoie BR, Wheeler BS, Ralston EJ, et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature. 2015;523:240–4.

109. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res. 2016;44:e70.

110. Oluwadare O, Cheng J. ClusterTAD: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data. BMC Bioinformatics. 2017;18:480.

111. Haddad N, Vaillant C, Jost D. IC-Finder: inferring robustly the hierarchical organization of chromatin folding. Nucleic Acids Res. 2017;45:e81.

112. Serra F, Baù 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:e1005665.

113. Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. Bioinformatics. 2016;32:1601–9.

114. Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics. 2014;30:i386–92.

115. Yan K-K, Lou S, Gerstein M. MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. PLoS Comput Biol. 2017;13:e1005647.

116. Norton HK, Emerson DJ, Huang H, Kim J, Titus KR, Gu S, et al. Detecting hierarchical genome folding with network modularity. Nat Methods. 2018;15:119–22.

117. Forcato M, Nicoletti C, Pal K, Livi C, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. Nat Methods. 2017;14:679–85.

118. Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test. Nat Commun. 2017;8:535.

119. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015;518:331–6.

120. Dong Q, Li N, Li X, Yuan Z, Xie D, Wang X, et al. Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice. Plant J. Wiley Online Library; 2018;94:1141–56.

121. Zhao Y-T, Kwon DY, Johnson BS, Fasolino M, Lamonica JM, Kim YJ, et al. Long genes linked to autism spectrum disorders harbor broad enhancer-like chromatin domains. Genome Res. 2018;28:933–42.

122. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, et al. A three-dimensional model of the yeast genome. Nature. 2010;465:363–7.

123. Ay F, Bailey TL, Noble W. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. 2014;24:999–1011.

124. Kaul A, Bhattacharyya S, Ay F. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. Nat Protoc. 2020;15:991–1012.

125. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38:576–89.

126. 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:e0174744.

127. Bhattacharyya S, Chandra V, Vijayanand P, Ay F. Identification of significant chromatin contacts from HiChIP data by FitHiChIP. Nat Commun. 2019;10:4221.

128. Hwang Y-C, Lin C-F, Valladares O, Malamon J, Kuksa PP, Zheng Q, et al. HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. Bioinformatics. 2015;31:1290–2.

129. Carty M, Zamparo L, Sahin M, González A, Pelossof R, Elemento O, et al. An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. Nat Commun. 2017;8:ncomms15454.

130. Xu Z, Zhang G, Jin F, Chen M, Furey TS, Sullivan PF, et al. A hidden Markov random field-based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data. Bioinformatics. 2016;32:650–6.

131. Xu Z, Zhang G, Wu C, Li Y, Hu M. FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. Bioinformatics. 2016;32:2692–5.

132. Alinejad-Rokny H, Ghavami R, Rabiee HR, Rezaei N. MaxHiC: robust estimation of chromatin interaction frequency in Hi-C and capture Hi-C experiments. bioRxiv [Internet]. biorxiv.org; 2020; Available from: https://www.biorxiv.org/content/10.1101/2020.04.23.056226v1.abstract

133. Cairns J, Freire-Pritchett P, Wingett SW, Várnai C, Dimond A, Plagnol V, et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. 2016;17:127.

134. 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;20:102.

135. Cao Y, Chen Z, Chen X, Ai D, Chen G, McDermott J, et al. Accurate loop calling for 3D genomic data with cLoops. Bioinformatics. 2020;36:666–75.

136. Vian L, Pękowska A, Rao SSP, Kieffer-Kwon K-R, Jung S, Baranello L, et al. The Energetics and Physiological Impact of Cohesin Extrusion. Cell. 2018;175:292–4.

137. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

138. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

139. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization [Internet]. arXiv [cs.LG]. 2014. Available from: http://arxiv.org/abs/1412.6980

140. Genovese CR, Roeder K, Wasserman L. False discovery control with p-value weighting. Biometrika. Oxford Academic; 2006;93:509–24.

141. Lareau CA, Aryee MJ. hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. Nat. Methods. 2018. p. 155–6.

142. Hong H, Jiang S, Li H, Du G, Sun Y, Tao H, et al. DeepHiC: A Generative Adversarial Network for Enhancing Hi-C Data Resolution. PLoS Comput Biol. 2020;16:e1007287.

143. Zhang Y, An L, Xu J, Zhang B, Zheng WJ, Hu M, et al. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. Nat Commun. 2018;9:750.

144. Liu T, Wang Z. HiCNN: a very deep convolutional neural network to better enhance the resolution of Hi-C data. Bioinformatics. 2019;35:4222–8.

145. Ester M, Kriegel H-P, Sander J, Xu X, Others. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996. p. 226–31.

146. Kraft K, Magg A, Heinrich V, Riemenschneider C, Schöpflin R, Markowski J, et al. Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. Nat Cell Biol. 2019;21:305–10.

147. Royer-Pokora B, Kunkel LM, Monaco AP, Goff SC, Newburger PE, Baehner RL, et

al. Cloning the gene for an inherited human disorder—chronic granulomatous disease—on the basis of its chromosomal location. Nature. Nature Publishing Group; 1986;322:32–8.

148. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, et al. Identification of the cystic fibrosis gene: genetic analysis. Science. 1989;245:1073–80.

149. Strathdee CA, Gavish H, Shannon WR, Buchwald M. Cloning of cDNAs for Fanconi's anaemia by functional complementation. Nature. 1992;358:434.

150. Wolffe A. Chromatin: Structure and Function. Academic Press; 1998.

151. Woodcock CL, Dimitrov S. Higher-order structure of chromatin and chromosomes. Curr Opin Genet Dev. 2001;11:130–5.

152. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012;148:84–98.

153. Beagrie RA, Scialdone A, Schueler M, Kraemer DCA, Chotalia M, Xie SQ, et al. Complex multi-enhancer contacts captured by genome architecture mapping. Nature. 2017;543:519–24.

154. Rubin AJ, Barajas BC, Furlan-Magaril M, Lopez-Pajares V, Mumbach MR, Howard I, et al. Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. Nat Genet. 2017;49:1522–8.

155. Montefiori LE, Sobreira DR, Sakabe NJ, Aneas I, Joslin AC, Hansen GT, et al. A promoter interaction map for cardiovascular disease genetics. Elife [Internet]. 2018;7. Available from: http://dx.doi.org/10.7554/eLife.35788

156. Chen H, Xiao J, Shao T, Wang L, Bai J, Lin X, et al. Landscape of Enhancer-Enhancer Cooperative Regulation during Human Cardiac Commitment. Mol Ther Nucleic Acids. 2019;17:840–51.

157. 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;51:1442–9.

158. Lu L, Liu X, Huang W-K, Giusti-Rodríguez P, Cui J, Zhang S, et al. Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases. Mol Cell. 2020;79:521–34.e15.

159. Qin Y, Grimm SA, Roberts JD, Chrysovergis K, Wade PA. Alterations in promoter interaction landscape and transcriptional network underlying metabolic adaptation to diet. Nat Commun. 2020;11:962.

160. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. 2015;16:22.

161. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol.

nature.com; 2010;28:1045–8.

162. Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, et al. BLUEPRINT to decode the epigenetic signature written in blood. Nat Biotechnol. nature.com; 2012;30:224–6.

163. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

164. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014;515:355–64.

165. Bai X, Shi S, Ai B, Jiang Y, Liu Y, Han X, et al. ENdb: a manually curated database of experimentally supported enhancers for human and mouse. Nucleic Acids Res. 2020;48:D51–7.

166. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

167. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017;12:2478–92.

168. Lin CY, Erkek S, Tong Y, Yin L, Federation AJ, Zapatka M, et al. Active medulloblastoma enhancers reveal subgroup-specific cellular origins. Nature. 2016;530:57–62.

169. Ron G, Globerson Y, Moran D, Kaplan T. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nat Commun. 2017;8:2237.

170. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal lari R, et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 2014;24:1–13.

171. Yao L, Shen H, Laird PW, Farnham PJ, Berman BP. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. Genome Biol. 2015;16:105.

172. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol Cell. 2018;71:858–71.e8.

173. Mehdi T, Bailey SD, Guilhamon P, Lupien M. C3D: a tool to predict 3D genomic interactions between cis-regulatory elements. Bioinformatics. 2019;35:877–9.

174. Zeng W, Wu M, Jiang R. Prediction of enhancer-promoter interactions via natural language processing. BMC Genomics. 2018;19:84.

175. Tao H, Li H, Xu K, Hong H, Jiang S, Du G, et al. Computational methods for the prediction of chromatin interaction and organization using sequence and epigenomic profiles. Brief Bioinform [Internet]. 2021; Available from:

http://dx.doi.org/10.1093/bib/bbaa405

176. Pott S, Lieb JD. What are super-enhancers? Nat Genet. Nature Publishing Group; 2014;47:8–12.

177. Oudelaar AM, Davies JOJ, Hanssen LLP, Telenius JM, Schwessinger R, Liu Y, et al. Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. Nat Genet. 2018;50:1744–51.

178. 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;174:744–57.e24.

179. Sabari BR, Dall'Agnese A, Boija A, Klein IA, Coffey EL, Shrinivas K, et al. Coactivator condensation at super-enhancers links phase separation and gene control. Science [Internet]. 2018;361. Available from: http://dx.doi.org/10.1126/science.aar3958

180. Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA. A Phase Separation Model for Transcriptional Control. Cell. 2017;169:13–23.

181. Smith NC, Matthews JM. Mechanisms of DNA-binding specificity and functional gene regulation by transcription factors. Curr Opin Struct Biol. 2016;38:68–74.

182. Wang X, Cairns MJ, Yan J. Super-enhancers in transcriptional regulation and genome organization. Nucleic Acids Res. 2019;47:11481–96.

183. Hu Z, Tee W-W. Enhancers and chromatin structures: regulatory hubs in gene expression and diseases. Biosci Rep [Internet]. 2017;37. Available from: http://dx.doi.org/10.1042/BSR20160183

184. Lee B-K, Jang YJ, Kim M, LeBlanc L, Rhee C, Lee J, et al. Super-enhancer-guided mapping of regulatory networks controlling mouse trophoblast stem cells. Nat Commun. 2019;10:4749.

185. 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;485:381–5.

186. Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell. 2013;153:320–34.

187. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153:307–19.

188. Ing-Simmons E, Seitan VC, Faure AJ, Flicek P, Carroll T, Dekker J, et al. Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. Genome Res. 2015;25:504–13.

189. Gong Y, Lazaris C, Sakellaropoulos T, Lozano A, Kambadur P, Ntziachristos P, et al. Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. Nat Commun. 2018;9:542.

190. Zhu D-L, Chen X-F, Hu W-X, Dong S-S, Lu B-J, Rong Y, et al. Multiple functional variants at 13q14 risk locus for osteoporosis regulate RANKL expression through long-range super-enhancer: LONG-RANGE MODULATION OF RANKL EXPRESSION BY BMD VARIANTS AT 13q14.11. J Bone Miner Res. Wiley; 2018;33:1335–46.

191. Hu H, Miao Y-R, Jia L-H, Yu Q-Y, Zhang Q, Guo A-Y. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. 2019;47:D33–8.

192. Jin J, Tian F, Yang D-C, Meng Y-Q, Kong L, Luo J, et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 2017;45:D1040–5.

193. Yevshin I, Sharipov R, Kolmykov S, Kondrakhin Y, Kolpakov F. GTRD: a database on gene transcription regulation—2019 update. Nucleic Acids Res. Oxford Academic; 2019;47:D100–5.

194. Jiang Y, Qian F, Bai X, Liu Y, Wang Q, Ai B, et al. SEdb: a comprehensive human super-enhancer database. Nucleic Acids Res. 2019;47:D235–43.

195. Khan A, Zhang X. dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic Acids Res. 2016;44:D164–71.

196. Wei Y, Zhang S, Shang S, Zhang B, Li S, Wang X, et al. SEA: a super-enhancer archive. Nucleic Acids Res. 2016;44:D172–9.

197. Chen C, Zhou D, Gu Y, Wang C, Zhang M, Lin X, et al. SEA version 3.0: a comprehensive extension and update of the Super-Enhancer archive. Nucleic Acids Res. 2020;48:D198–203.

198. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106:9362–7.

199. Freedman ML, Monteiro ANA, Gayther SA, Coetzee GA, Risch A, Plass C, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011;43:513–8.

200. Giral H, Landmesser U, Kratzer A. Into the Wild: GWAS Exploration of Non-coding RNAs. Front Cardiovasc Med. 2018;5:181.

201. Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome Res. 2014;24:1854–68.

202. Martin P, McGovern A, Massey J, Schoenfelder S, Duffus K, Yarwood A, et al. Identifying Causal Genes at the Multiple Sclerosis Associated Region 6q23 Using Capture Hi-C. PLoS One. Public Library of Science; 2016;11:e0166923.

203. McGovern A, Schoenfelder S, Martin P, Massey J, Duffus K, Plant D, et al. Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. Genome Biol. 2016;17:212.

204. Baxter JS, Leavy OC, Dryden NH, Maguire S, Johnson N, Fedele V, et al. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. Nat Commun. 2018;9:1028.

205. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42:D1001–6.

206. Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong C, Fumis L, Karamanis N, et al. Open Targets Platform: new developments and updates two years on. Nucleic Acids Res. 2019;47:D1056–65.

207. Beck T, Shorter T, Brookes AJ. GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. Nucleic Acids Res. 2020;48:D933–40.

208. Tian D, Wang P, Tang B, Teng X, Li C, Liu X, et al. GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals. Nucleic Acids Res. 2020;48:D927–32.

209. Li MJ, Liu Z, Wang P, Wong MP, Nelson MR, Kocher J-PA, et al. GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res. 2016;44:D869–76.

210. Doerge RW. Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet. 2002;3:43–52.

211. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.

212. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, et al. Identification of genetic variants that affect histone modifications in human cells. Science. 2013;342:747–9.

213. Grubert F, Zaugg JB, Kasowski M, Ursu O, Spacek DV, Martin AR, et al. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell. 2015;162:1051–65.

214. Smith AK, Kilaru V, Kocak M, Almli LM, Mercer KB, Ressler KJ, et al. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. BMC Genomics. 2014;15:145.

215. Nicodemus-Johnson J, Myers RA, Sakabe NJ, Sobreira DR, Hogarth DK, Naureckas ET, et al. DNA methylation in lung cells is associated with asthma endotypes and genetic risk. JCI Insight. 2016;1:e90151.

216. Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat Genet. 2018;50:1140–50.

217. Greenwald WW, Chiou J, Yan J, Qiu Y, Dai N, Wang A, et al. Pancreatic islet chromatin accessibility and conformation defines distal enhancer networks of type 2 diabetes risk. BioRxiv. 2018;299388.

218. Yu J, Hu M, Li C. Joint analyses of multi-tissue Hi-C and eQTL data demonstrate close spatial proximity between eQTLs and their target genes. BMC Genet. 2019;20:43.

219. Xia K, Shabalin AA, Huang S, Madar V, Zhou Y-H, Wang W, et al. seeQTL: a searchable database for human eQTLs. Bioinformatics. 2012;28:451–2.

220. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res. 2016;44:D877–81.

221. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45:1238–43.

222. Gong J, Wan H, Mei S, Ruan H, Zhang Z, Liu C, et al. Pancan-meQTL: a database to systematically evaluate the effects of genetic variants on methylation in human cancer. Nucleic Acids Res. 2019;47:D1066–72.

223. Zheng Z, Huang D, Wang J, Zhao K, Zhou Y, Guo Z, et al. QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. Nucleic Acids Res. 2020;48:D983–91.

224. Shumway M, Cochrane G, Sugawara H. Archiving next generation sequencing data. Nucleic Acids Res. 2010;38:D870–1.

225. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, et al. The European Nucleotide Archive. Nucleic Acids Res. 2011;39:D28–31.

226. Nozaki T, Imai R, Tanbo M, Nagashima R, Tamura S, Tani T, et al. Dynamic Organization of Chromatin Domains Revealed by Super-Resolution Live-Cell Imaging. Mol Cell. 2017;67:282–93.e7.

227. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014;157:1262–78.

228. Doudna JA, Charpentier E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science. 2014;346:1258096.

Chapter 2 Methods

Development of Hi-C data analysis components for accurate chromatin interaction profiles

Development of Hi-C data analysis components for accurate chromatin interaction profiles

Abstract

High resolution chromosome conformation capture sequencing (Hi-C) is the state-of-the-art sequencing technique used to identify three-dimensional chromatin structure inside the nucleus. Due to the complex nature of Hi-C data and different laboratory protocols available, reproducible computational methods are essential to construct genome-wide chromatin contact maps and identify informative chromatin interactions. There are currently a number of standard workflows designed for analysing Hi-C data, however quality control metrics to determine the quality of a Hi-C library are difficult to obtain, as is the ability to visualise Hi-C interactions with other genomics data. In this methods chapter, I develop a number of tools and analysis strategies designed to improve the analysis of all Hi-C data types and enable many of the analyses developed in Chapter 3 and 4, focusing specifically on three components; Quality Control, Genome Alignment and Visualisation. For quality control, I developed *HiC-QC* to conveniently summarise important metrics required for assessing the quality of a Hi-C library. For genome alignment, I investigate the appropriate usage of two popular aligners (Bowtie2 and BWA) in their performance in aligning Hi-C sequencing reads to a reference genome. Lastly, two visualisation methods, the *HiC-integrationmap* and the *integration-tracks* plot were also developed to facilitate the visualisation of integration between Hi-C interactions and other

genomics data. These analyses and tools aim to give researchers a better guide to accounting for limitations in Hi-C datasets and to enable the reuse of these datasets for functional investigations.

Background

High-resolution chromosome conformation capture sequencing (Hi-C) assay (Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014) have been widely used in studies to reveal the 3D structure of the genome (Dixon *et al.*, 2012, 2015; Jin *et al.*, 2013) and investigate the gene regulation contributed by the chromatin interactions (Mifsud *et al.*, 2015; Javierre *et al.*, 2016). However, Hi-C sequencing data is complex because complicated experimental procedures such as cross-linked DNA digestion and ligation can introduce uninformative data such as self-ligating DNA or random ligation during the construction of the Hi-C library, hindering the identification of genuine chromatin interactions (Ay, Bailey and Noble, 2014; Mifsud *et al.*, 2017). Therefore, it is important to learn the computational procedures to analyse Hi-C data properly to identify chromatin interactions and construct Hi-C contact maps (Lajoie, Dekker and Kaplan, 2015).

The analysis of Hi-C data generally comprises four fundamental steps; alignment of sequencing reads, filtering to identify informative interaction pairs, generating an interaction matrix and normalization of the interaction matrix (Figure 1). However, prior to this analysis, data preparation and quality control stages must be conducted, and is often overlooked that is critical to downstream analysis.

Raw sequencing data from a Hi-C experiment is similar to that of other sequencing data (e.g. RNA sequencing or chromatin immunoprecipitation sequencing) in that it is necessary to conduct quality control processes in order to evaluate sequencing quality. In an RNA-seq experiment for example, quality control is achieved by the inspection of metrics such as sequencing base quality score, GC content, sequence duplication levels and overrepresented sequence quantification (Figure 1). QC metrics are often summarised using tools such as FastQC (Andrews and Others, 2010) for a single dataset, and multiQC (Ewels *et al.*, 2016) or ngsReports (Ward, To and Pederson, 2020) for multiple datasets.

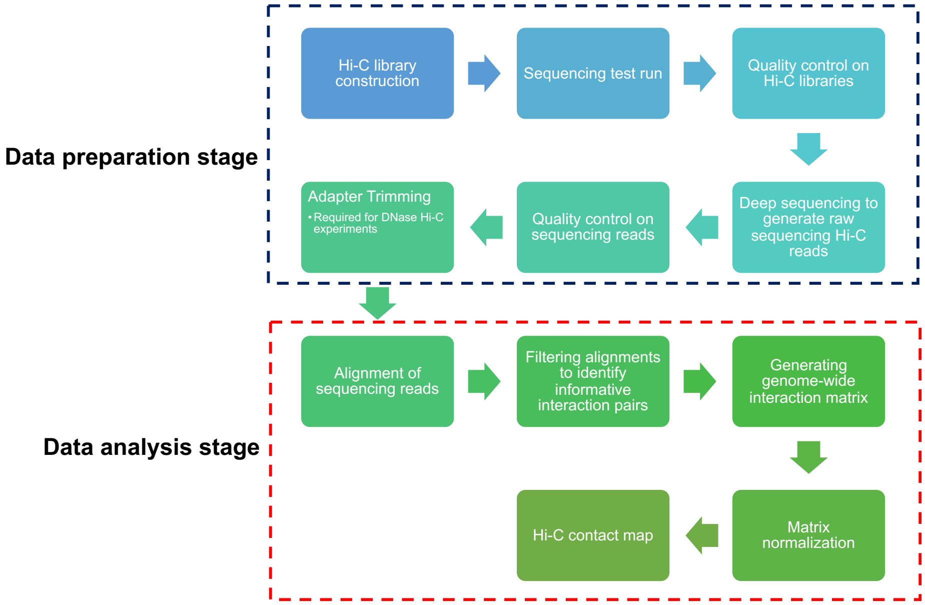

Data preparation stage
Hi-C library construction
Sequencing test run
Quality control on Hi-C libraries
Deep sequencing to generate raw sequencing Hi-C reads
Quality control on sequencing reads
Adapter Trimming
• Required for DNase Hi-C experiments
Data analysis stage
Alignment of sequencing reads
Filtering alignments to identify informative interaction pairs
Generating genome-wide interaction matrix
Matrix normalization
Hi-C contact map

Figure 1: Comprehensive flow-diagram of a Hi-C experiment. The blue box indicates the Hi-C data preparation stage, while the red box indicates the data analysis stage of Hi-C data.

Hi-C sequencing data is different to that of other sequencing data in a number of important ways. Firstly, generation of a high resolution, genome-wide Hi-C contact map requires higher sequencing depth than that required for many sequencing applications. A depth of ~7 billion raw reads was required to reach a resolution of 950 bp in the construction of a contact map of the GM12878 cell line (Rao *et al.*, 2014), which is far larger than a standard 30x whole genome sequencing sample (Rieber *et al.*, 2013). Secondly, the protocol for construction of a Hi-C library is complicated by DNA crosslinking, digestion and ligation steps (Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014). The failure of any of these steps will result in a Hi-C library with poor quality. To reduce the risk of conducting downstream analysis of poor quality Hi-C libraries it is essential to conduct quality control to ensure that libraries pass metrics for both quality and sequencing depth (Figure 1) (Patel and Jain, 2012; Rao *et al.*, 2014). Additionally, whilst Hi-C libraries usually do not produce a significant amount of adapter contamination, certain digestion steps, especially in DNase Hi-C protocol, can generate DNA fragments that retain both adapter sequences and Illumina primers (Ma *et al.*, 2018). In the event of this type of adapter contamination, it is important that the sequencing adapters are removed to ensure better mapping results (Figure 1), with tools such as AdapterRemoval (Schubert, Lindgreen and Orlando, 2016) or cutadapt (Martin, 2011) being commonly used for this purpose.

After data preparation, the data analysis stage is firstly started by linking the sequencing data information with a genome location via alignment (Figure 1). Due to the different primary location of each DNA read fragment, the alignment of

Hi-C paired-end reads directly to a reference genome should be avoided (Figure 2). Instead, one should treat Hi-C paired-end reads as two single-end reads, align them separately to the genome, and then match them by their unique read pair identification (Lieberman-Aiden *et al.*, 2009). A number of mapping strategies have been developed specifically for alignment of Hi-C data (Table 1). One such aligner in common use is the Burrows-Wheeler Aligner (BWA) software. The BWA-MEM algorithm employs a “-5SPM" flag to disable read pairing specifically for use in mapping Hi-C data (Li, 2013).

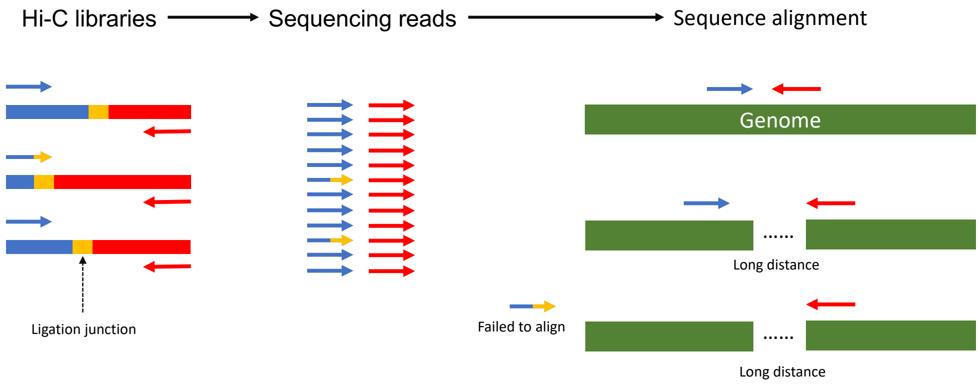


Figure 2: Schematic figure of the alignment process of Hi-C data. Each pair of blue and red arrows indicates Hi-C sequencing read pairs. The yellow squares indicate the ligation junctions, which result from the successful ligation step during constructing the Hi-C library. The yellow arrows indicate the ligation junction contaminations during sequencing.

Table1: Software tools to map, filter, quantify and normalise Hi-C data.

Standard analysis of Hi-C data			
Procedures	Softwares/Strategies	Reference	Source
Mapping	BWA-MEM -5SPM	(Li, 2013)	https://github.com/lh3/bwa
	Pre-truncation	(Wingett *et al.*, 2015)	https://www.bioinformatics.babraham.ac.uk/projects/hicup/
	Post-truncation	(Servant *et al.*, 2015)	https://github.com/nservant/HiC-Pro
	Iterative mapping	(Imakaev *et al.*, 2012)	https://github.com/mirnylab/hiclib-legacy
Filtering	Low mapping quality reads	NA	NA
	Non-unique alignments		
	Singletons		
	Dangling ends pair		
	Duplication		
Quantification	Binning	(Lieberman-Aiden *et al.*, 2009)	NA
Normalization(Explicit)	Yaffe and Tanay's model	(Yaffe and Tanay, 2011)	NA
	HiCNorm	(Hu *et al.*, 2012)	https://github.com/ren-lab/HiCNorm
	Jin's model	(Jin *et al.*, 2013)	NA
	OneD	(Vidal *et al.*, 2018)	https://github.com/qenvio/dryhic
Normalization(Implicit)	Vanilla coverage	(Lieberman-Aiden *et al.*, 2009)	NA
	ICE	(Imakaev *et al.*, 2012)	https://github.com/nservant/HiC-Pro
	Knight and Ruiz's model	(Rao *et al.*, 2014)	https://github.com/aidenlab/juicer

Aligning reads separately to the genome is the basic rule of Hi-C alignment (Lieberman-Aiden *et al.*, 2009). However, mapping of Hi-C reads is not straightforward given the presence of ligation junction sequence within a read can lead to mapping failure (Figure 2). Furthermore, it is not possible to know *a* *priori* whether the ligation junction will be found within either the forward, reverse or both ends of the sequence reads. In order to overcome this issue two ligation junction targeting strategies have been developed (Table 1). These include a truncation method before mapping (Wingett *et al.*, 2015), which trims off the sequence downstream of the ligation junction before mapping, and an unmapped read rescuing strategy after mapping (Servant *et al.*, 2015), which trims off the sequence downstream of the ligation junction in unmapped reads, and is then followed by a second round of mapping using the trimmed sequences. In addition an iterative mapping strategy may address the ligation junction issue by first trimming the reads into short sequences (25 bp) which are then extended by 5 bp to the reads iteratively until a unique alignment or maximum read length is achieved (Imakaev *et al.*, 2012).

After alignment, the next step is to remove low mapping quality reads and non-unique alignments (Figure 1 and Table 1) using tools such as SAMTools (Li *et al.*, 2009). Subsequent to this filtering, uninformative read pairs are identified based on their unique features and removed. Reads that fail to match with a mate-pair (i.e. singletons) are removed, as they fail to infer a valid interaction (Servant *et al.*, 2015). When restriction enzyme digestion is used within the HiC library preparation, we expect each read of a pair to be observed in different restriction fragments, i.e. DNA segments between every two restriction sites. If both reads of a pair are located in the same fragment, this would indicate either self-ligation products of the fragment or ‘dangling ends’ (Lajoie, Dekker and

Kaplan, 2015; Servant *et al.*, 2015). These fragments also need to be removed as they introduce a known bias from library preparation. Finally, PCR duplicates are removed since they will overinflate the quantification of Hi-C contacts (Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014), through identification and removal of read pairs that share identical 5' and 3' ends.

After mapping QC filtering is complete, the remaining read pairs are considered valid 3D interactions. In order to quantify the interactions (Figure 1), an interaction frequency matrix is generated by binning the genome into equal size bins and then mapping the valid interaction pairs to each individual bin (Lieberman-Aiden *et al.*, 2009). The choice of bin size depends on the number of identified valid interaction pairs, which is a function of the coverage and quality of the sequencing data, and it is usually used as a reference of the resolution of Hi-C dataset (Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014). In short, the smaller the bin size, the higher the resolution of detected interactions. It is important to choose a suitable bin size for Hi-C data, as mapping to undersized bins will result in a sparse, zero-inflated matrix, which impacts the subsequent normalisation step and downstream analysis including identifying statistically significant interactions and structural domains such as topologically-associated domains (TADs) or A/B compartments. The smallest bin size (and therefore highest resolution) seen in human Hi-C studies has been one kilobase bin size, and required approximately 4.9 billion interactions (Rao *et al.*, 2014). Most studies however, are analysing data at lower resolutions, with a range of 10-40 kb bin size (Javierre *et al.*, 2016; Taberlay *et al.*, 2016; Rodrigues *et al.*, 2018; Barisic *et al.*, 2019).

The selection of bin size should also take into consideration the biological question at hand. For example, if a study aims to refine a region to be investigated based on the 3D structure, larger bin size such as 50-150 kb can be used to identify TADs or A/B compartments at multi-kilobases to megabases scale; while smaller bin sizes, such as 5-10 kb allow for the investigation of long-distance interactions impacting regulatory elements such as transcription factors or enhancers (Dixon *et al.*, 2012; Rao *et al.*, 2014).

After the creation of a interaction matrix, further normalisation is required to account for biases introduced in the Hi-C library preparation (Figure 1) (Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014; Lajoie, Dekker and Kaplan, 2015; Servant *et al.*, 2015). A number of normalisation methods are widely used (Table 1) and can be classified into explicit and implicit methods based on the assumption of their models (Pal, Forcato and Ferrari, 2019). Explicit normalisation methods are designed to directly account for the sources of bias known to have originated from the Hi-C protocols or sequencing technologies (Yaffe and Tanay, 2011). There are three major sources of bias that are built into a probabilistic background model: non-specific digestion events in Hi-C experiment with restriction enzymes, restriction fragment lengths bias and GC content of the regions near the ligation sites (Yaffe and Tanay, 2011). By contrast, implicit normalisation approaches are designed to normalize Hi-C data without making assumptions regarding the source of biases in the experiment, all of which are based on a matrix balancing method developed by Sinkhorn and Knopp (Sinkhorn and Knopp, 1967). One of the most popular methods is Iterative correction and eigenvector decomposition (ICE). This method was developed based on the assumption that the total number of interaction signals will be the same across all genomic loci (Imakaev *et al.*, 2012). Recently, a new matrix balancing algorithm (Knight and Ruiz, 2013), which performs much faster with similar convergence properties as Sinkhorn-Knopp's method, has been used to normalize interaction matrices in ultra-high-resolution Hi-C datasets (1 kb resolution) (Rao *et al.*, 2014).

Protocols/workflows for the analysis of Hi-C data have been optimised and standardised in computational pipelines, published studies and research projects including HiC-Pro, Juicer and the 4DN project (Servant *et al.*, 2015; Durand *et al.*, 2016; Dekker *et al.*, 2017). However, there are three critical aspects of the analysis that require further optimization and investigation. Firstly, while Juicer and HiC-Pro, two of the most popular pipelines in current use, both summarise statistics at each step of the pipeline, there are no tools that conveniently summarise and import statistics metrics into a single document to assist QC on Hi-C libraries. Secondly, the choices of aligner and alignment strategy are different from study to study, mainly driven by the usage of developed pipelines, the types of protocol to generate Hi-C library and the quality of sequencing data.

HiC-Pro for example was developed based on the Bowtie2 aligner rather than the BWA-MEM aligner due to BWA-MEM being incapable of performing split read analysis, where each read in a read pair was considered as single-end sequencing data and this is essential for Hi-C data mapping. However in the latter updates of BWA-MEM, parameters that were specifically designed for Hi-C data mapping (i.e. the '5SPM' flag) were included in the program, enabling the split parameters. This program benefits from having a shorter running time than Bowtie2, making it a preferable choice for alignment in later-published pipelines, such as Juicer. Lastly, the visualisation methods of Hi-C data, such as contact probability heatmap, circular plot, local arc track and multi-track visualisation (Akdemir and Chin, 2015; Kerpedjiev *et al.*, 2018) are limited, particularly when integration of multiple genomic data types is considered.

In this chapter, in order to optimize current computational methods of analysing Hi-C data, we first developed HiC-QC, a tool to summarise the metrics to assist quality control of Hi-C libraries. We then conduct a systematic comparison between two aligners Bowtie2 and BWA with different Hi-C data. Finally, we developed HiC-integrationmap and integration-tracks plot to facilitate the visualisation of Hi-C interaction integration, providing a new and complimentary tool for Hi-C data analysis. The source code of all developed tools are publicly available at https://github.com/ningbioinfostruggling/HiCvisualisation.

Results

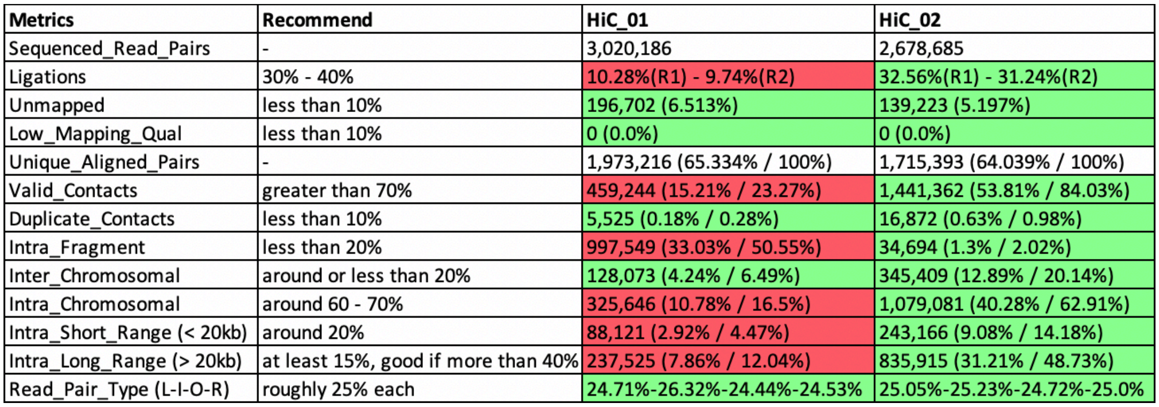
Quality control of Hi-C libraries using HiC-QC

Quality control (QC) in Hi-C analysis is an essential first step in determining the quality of a sequencing library. Constructing a comprehensive chromatin contact map is expensive because it requires high sequencing depth (Rao *et al.*, 2014). It is a risk therefore to conduct deep sequencing without first assessing the quality of the Hi-C library preparation. In previous Hi-C studies (Oksuz *et al.*, 2020; Rao *et al.*, 2014; Belaghzal, Dekker and Gibcus, 2017), it is common practice for the Hi-C libraries to be first sequenced at a low sequencing depth, generating 1~3 million reads, before undergoing the standard downstream analysis steps (Figure 1). QC is then conducted by obtaining statistics from each step to evaluate the quality of the Hi-C library. Finally, libraries of good quality are chosen to be sequenced to a depth of hundreds of millions of reads to be used for answering biological questions or constructing chromatin contact maps. To ensure the quality of Hi-C libraries, based on the analysis procedures of the HiC-Pro pipeline, here we developed a computational tool called HiC-QC to summarise useful statistics in the QC of Hi-C libraries.

While HiC-QC does not generate its own statistics, it relies on statistical metrics from other programs, such as samtools, to summarise 13 different metrics throughout the Hi-C analysis process. These statistics include many of the recommended statistics from published studies, such as the uniquely mapping rate, the duplication rate and the ratio of inter- and intra- chromosomal interactions. Standard sequencing and mapping statistics such as “Sequenced Read Pairs”, “Unmapped”, “Low Mapping Quality” and “Unique Aligned Pairs”, all of which are common in QC of other types of sequencing data, are included (Figure 3). The “Unmapped” flag indicates the number of reads that failed to align to the reference genome. It has been suggested that a mapping rate below 90% (i.e. more than 10% reads unmapped) indicates that either the sequencing run is problematic or that the sample was contaminated (Lieberman-Aiden *et al.*, 2009; Jin *et al.*, 2013; Rao *et al.*, 2014). The “Ligation” flag represents a statistic that is specifically for Hi-C data generated by restriction enzymes digested protocols. It reveals the number of sequencing reads that contain the ligating DNA sequence, which is introduced when constructing the Hi-C library.

The ligation sequence is dependent on which restriction enzyme was used in the experiment, for example, 6-base cutters that generate one-base overhang result in ligation sequences that are found at both ends, such as the ligation sequence of HindIII (cutting sequence A^AGCTT) which will result in a AAGCTAGCTT ligation sequence. On the other hand, the ligation sequences of 4-base cutters are generated by their digest sequences, the ligation sequence of MboI (cutting sequence ^GATC) for example will result in a GATCGATC ligation sequence. The number of sequencing reads that include the ligation sequence can be used for estimating if the ligation step is successful. Although it depends on the insert size and read length, it was suggested that the percentage of raw sequencing reads that contain ligation sequence should be around 30% to 40% for a successful experiment (300-500bp insert size and 101bp read length) (Rao *et al.*, 2014; Servant *et al.*, 2015). The “Valid contact” flag indicates the number of reads that remain following the filtering step, and are hence regarded as valid interaction pairs. The “Duplicate contacts” flag represents interaction pairs that share the exact same sequenced DNA. These are typically duplicated DNA fragments generated by PCR during sequencing and therefore the amount of them can reflect the quality of the sequencing run.

Besides duplicated pairs, another type of uninformative pair are both forward and reverse reads mapped to the same restriction fragments, which we call “Intra-fragment pairs”. It has been suggested that a percentage of intra-fragment pairs greater than 20% would indicate the failure of either the digestion or ligation steps during the experiment (Rao *et al.*, 2014; Servant *et al.*, 2015). One important feature of the identified Hi-C contacts is the distance between two anchors of any interactions. Using the distance between anchors the contacts can be classified into short-range (<= 20 kb) intra-chromosomal interactions, long-range (> 20 kb) intra-chromosomal interactions and inter-chromosomal interactions (Figure 3). It has been suggested that the percentage of long-range intra-chromosomal interactions reveals the quality of a Hi-C library, with a score of 40% long-range unique contacts being considered a good candidate for further deep sequencing (Rao *et al.*, 2014). As a final quality control the distribution of the read pair direction of intra-chromosomal contacts should be approximately even for a high quality Hi-C experiment, with a skewed distribution indicating that the observed interactions are not the result of a close proximity ligation (Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014; Servant *et al.*, 2015; Durand *et al.*, 2016).



Metrics	Recommend	HiC_01	HiC_02
Sequenced_Read_Pairs	-	3,020,186	2,678,685
Ligations	30% - 40%	10.28%(R1) - 9.74%(R2)	32.56%(R1) - 31.24%(R2)
Unmapped	less than 10%	196,702 (6.513%)	139,223 (5.197%)
Low_Mapping_Qual	less than 10%	0 (0.0%)	0 (0.0%)
Unique_Aligned_Pairs	-	1,973,216 (65.334% / 100%)	1,715,393 (64.039% / 100%)
Valid_Contacts	greater than 70%	459,244 (15.21% / 23.27%)	1,441,362 (53.81% / 84.03%)
Duplicate_Contacts	less than 10%	5,525 (0.18% / 0.28%)	16,872 (0.63% / 0.98%)
Intra_Fragment	less than 20%	997,549 (33.03% / 50.55%)	34,694 (1.3% / 2.02%)
Inter_Chromosomal	around or less than 20%	128,073 (4.24% / 6.49%)	345,409 (12.89% / 20.14%)
Intra_Chromosomal	around 60 - 70%	325,646 (10.78% / 16.5%)	1,079,081 (40.28% / 62.91%)
Intra_Short_Range (< 20kb)	around 20%	88,121 (2.92% / 4.47%)	243,166 (9.08% / 14.18%)
Intra_Long_Range (> 20kb)	at least 15%, good if more than 40%	237,525 (7.86% / 12.04%)	835,915 (31.21% / 48.73%)
Read_Pair_Type (L-I-O-R)	roughly 25% each	24.71%-26.32%-24.44%-24.53%	25.05%-25.23%-24.72%-25.0%

Figure 3: An example of output from HiC-QC. Statistics that can be used for quality control of a Hi-C dataset is summarised by HiC-QC. When output as a Microsoft Excel spreadsheet statistics for a library are coloured to indicate general quality. Red indicates statistics that fall outside recommended ranges, green within.

HiC-QC is able to search through the output directory created by HiC-Pro (Servant *et al.*, 2015) to obtain the above summary statistics from each step and output a summary in either comma-separated values (csv) format or Microsoft Excel format. These two formats allow a user with command line experience to interact with the data directly through a command line interface (csv), or through Excel for users more familiar with the Microsoft interface. An additional feature of the Microsoft Excel format output is the conditional colouration of each cell (coloured red or green) indicating if the statistics fall within recommended ranges for high quality data. Figure 3 gives an example of the HiC-QC output in Excel format. The second library, HiC_02, achieves quality metrics within recommended ranges for all assessed statistics, indicating this library is a good candidate for deep sequencing. However only 31.22% of the unique alignments of the first library HiC_01 were identified as valid Hi-C interactions, and 50.62% of the alignments are intra-fragment pairs. Together with the statistics that only around 10% of the raw sequencing reads contain the ligation sequence, we can speculate that HiC_01 may have failed at the ligation step during library construction.

Comparing alignment strategies of Hi-C data

The 3D conformation of a chromosome consists of numerous distal physical interactions between DNA from different regions. Ideally, each DNA fragment in the Hi-C library originates from two DNA fragments that are cross-linked and may be located in different areas across the genome (Figure 2). While HiC-QC can facilitate the quality control step of the analysis, it is difficult to select a suitable aligner to map Hi-C data to the reference genome. Two widely used choices for aligning Hi-C data to the reference genome in common use are BWA and Bowtie2 (Table 2). However, there is no study that systematically compares these two aligners regarding their ability to map Hi-C data generated from different protocols, such as *in situ* Hi-C, capture Hi-C and DNase Hi-C.

Table 2: Aligner and mapping quality threshold choice of published Hi-C data analysis pipeline and studies.

Pipeline/*Study*	Aligner	MAPQ threshold	Extra information	Reference
Juicer	BWA	30	-	(Durand *et al.*, 2016)
HiC-Pro	Bowtie2	10	Post-trimmed reads if restriction sites appeared	(Servant *et al.*, 2015)
HiCUP	Bowtie2	30	Pre-trimmed reads if restriction sites appeared	(Wingett *et al.*, 2015)
HiCPipe	BWA	10	-	(Yang *et al.*, 2020)
FAN-C	BWA/Bowtie	3/30	Provide both aligners for user	(Kruse, Hug and Vaquerizas, 2020)
HIPPIE	BWA	30	-	(Hwang *et al.*, 2015)
diffHiC	Bowtie2	10	Pre-trimmed reads if restriction sites appeared	(Lun and Smyth, 2015)
in situ Hi-C	BWA	30	-	(Rao *et al.*, 2014)
Capture Hi-C	Bowtie2	30	Pre-trimmed reads if restriction sites appeared	(Mifsud *et al.*, 2015)
FIRE	BWA	10	-	(Schmitt *et al.*, 2016)
Compendium of PCHi-C	BWA	10	-	(Jung *et al.*, 2019)
DNase Hi-C	BWA	30	-	(Ramani *et al.*, 2016)

In order to compare Bowtie2 and BWA based on in their ability to correctly map different types of Hi-C data to the genome, we obtained three good quality datasets from published studies, including a *in situ* Hi-C library of Jurkat cell line (Lucic *et al.*, 2019), an *in situ* DNase Hi-C library of RUES2 cell line (Bertero *et al.*, 2019), and a Capture Hi-C library of GM12878 cell line (Jung *et al.*, 2019). Additionally, in order to establish how these aligners perform with low-mapping rates, we also include a mouse *in situ* Hi-C library of CH12-LX cell line (Rao *et al.*, 2014). Notably, the following experiments are mainly focused on comparing the capability of different aligners to map different types of Hi-C data. In order to account for the variability of the library size, percentage instead of count will be used in the following comparison analyses. It has previously been suggested that sequencing reads from DNase Hi-C data is impacted by Illumina sequencing adapters (Ma *et al.*, 2018), hence we decided to trim the sequencing adapters prior to the alignment process using AdapterRemoval (Schubert, Lindgreen and Orlando, 2016) across all datasets. Trimmed data was then aligned to the human genome using BWA-mem or Bowtie2 with strategies specifically designed for Hi-C data. In this comparison, we used hg19 genome as the reference genome because annotation databases including Roadmap Epigenomics (Roadmap Epigenomics Consortium et al., 2015), GTEx (GTEx Consortium et al., 2017) and the ENCODE project (ENCODE Project Consortium et al., 2020) are still mostly focusing on hg19 even though hg38 has been available since 2013. Additionally, hg19 is used in the tutorial of widely used HiC analysis workflows such as HiC-Pro (Servant et al., 2015) and Juicer (Durand et al., 2016). We acknowledge however, that recent new versions of the human genome, particularly ones with fully resolved Telomere-to-Telomere human chromosomes (Miga et al. 2020), would significantly benefit the analysis of Hi-C datasets, potentially identifying novel areas of function. Notably, since the pre-truncated strategy was popularly used with Bowtie2 in the published pipeline (Table 1), we used the pre-truncation function from HiCUP (Wingett *et al.*, 2015) to process sequencing reads before alignment with Bowtie2. BWA-mem was used with the parameter “SP5M”, which was specifically designed for Hi-C data processing. In order to investigate how these strategies impact alignment, we also include a comparison with two aligners mapping the *in situ* Hi-C data of the Jurkat cell line using default parameters. Finally, the aligned data were processed to identify interactions using Pairtools (https://github.com/mirnylab/pairtools) with a mapping quality threshold of 30.

After processing, we found that BWA tends to have a higher mapping rate compared to Bowtie2, even for the low-mapping rate dataset (CH12-LX) (Figure 4A). However, the mapped reads reported by Bowtie2 tend to have a high mapping quality (MAPQ 30), with 15% to 20% of the mapped reads in human data analysed by BWA being below the mapping quality threshold (Figure 4). Although the two aligners generated different alignment counts (alignment pair with MAPQ>=30), the identified interactions counts are similar for *in situ* Hi-C data (Jurkat) and Capture Hi-C data (GM12878) (Figure 4A). However, when looking at DNase Hi-C specifically, BWA generated more Hi-C interactions (1.86 fold), which was somewhat validated by the fact that most published research of DNase Hi-C data analyses tended to prefer BWA to conduct sequence alignment (Deng *et al.*, 2015; Ramani *et al.*, 2016; Bertero *et al.*, 2019). Additionally, when comparing Hi-C data-specific strategies with default settings of both aligners, we found that the mapping rate significantly decreased with default parameters (Figure 4A).

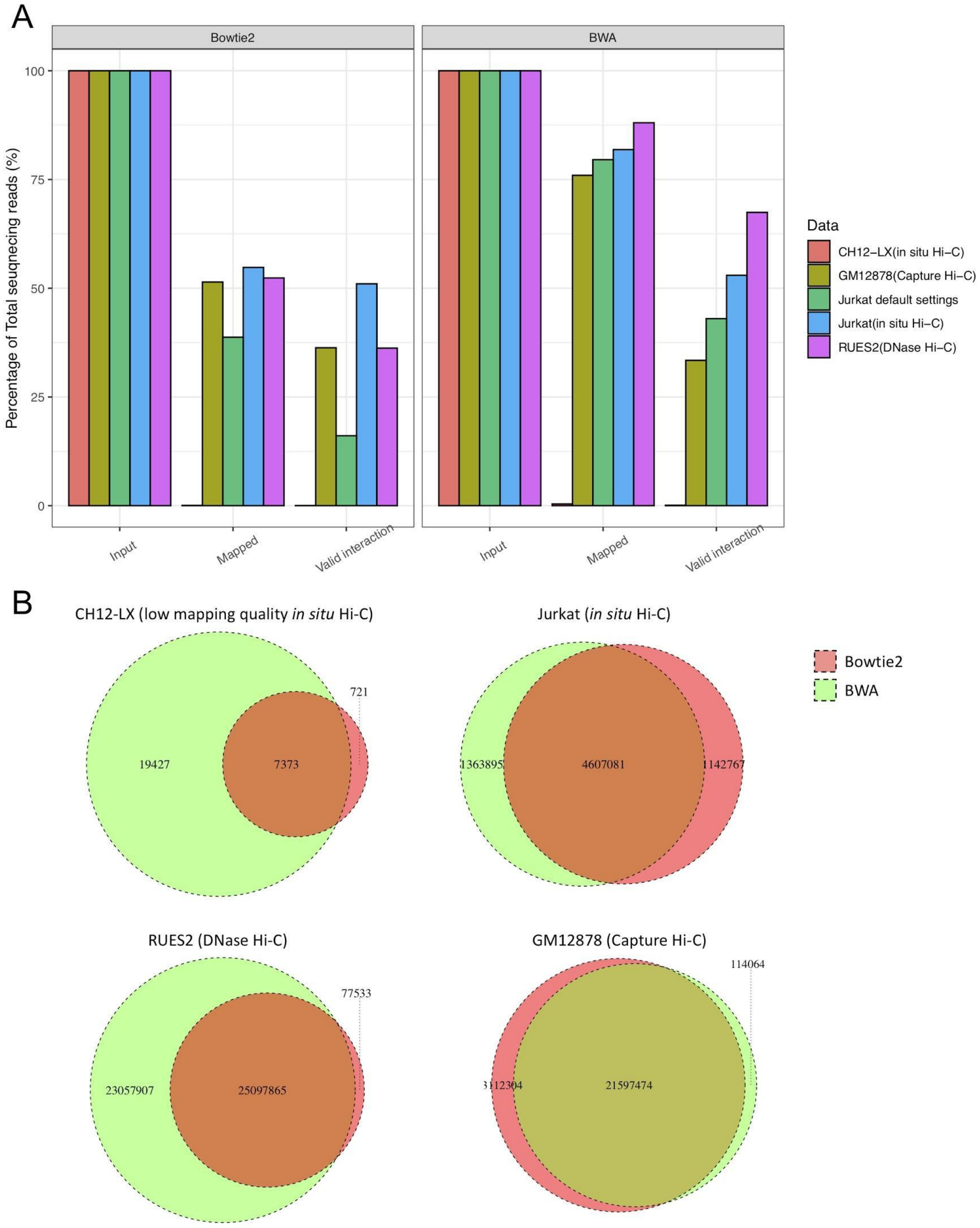


Figure 4: Aligner comparison for different types of Hi-C data. A: the fraction of sequencing reads that successfully mapped to the genome and be identified as

Hi-C interactions. B: intersection between the identified interactions of different types of Hi-C data processed by BWA and Bowtie2.

By investigating the identified interactions, we further observed that the same sample processed by two different aligners shared a high fraction of identified Hi-C interactions (Figure 4B). For in situ Hi-C data and Capture Hi-C data, on average 83.76% of the Bowtie2-mapped interactions are found in 82.28% of the BWA-mapped interactions. However, 91.09% and 99.69% of the Bowtie2-mapped interactions of the low-mapping rate data (CH12-LX) and DNase Hi-C data (RUES2) are found in BWA-mapped interactions, respectively, while 72.49% (CH12-LX) and 51.64% (RUES2) of the BWA-mapped interactions are not detected by Bowtie2-mapped data (Figure 4B).

In summary, for *in situ* Hi-C and capture Hi-C data, Bowtie2 and BWA seem to perform equally well with Hi-C data-specific parameters/strategy, while BWA requires strict filtering of uninformative alignments to identify Hi-C interactions. However, for low-mapping rate data and DNase Hi-C data, BWA outperforms Bowtie2 by identifying extra informative interactions.

Visualisation of Hi-C data integration with HiC-integrationmap and integration-tracks plot

One of the most popular approaches to visualize Hi-C contacts is plotting interaction intensity, which are represented by the read count of interacting bins, as a heatmap (Figure 5A). Additionally, a number of published studies have used arc plot (Javierre *et al.*, 2016; Jung *et al.*, 2019) or circos plot (Vieux-Rochas *et* *al.*, 2015; Klocko *et al.*, 2016) to obtain a better visualisation of Hi-C interactions in specific regions. However, it is difficult to make use of the Hi-C interactions by only visualising the contacts alone without epigenomic annotations. Therefore, one important downstream analysis of Hi-C data is to integrate with other epigenomics data to uncover the regulatory mechanism governed by 3D structure (Javierre *et al.*, 2016; Schmitt *et al.*, 2016; Liu *et al.*, 2020). Some interactive browsers, such as the UCSC browser (Kent et al., 2002) or the WashU Epigenome browser (Li et al., 2019) offer to overlay various types of sequencing data such as ATAC-seq, ChIP-seq, RNA-seq and SNPs data to integrate with Hi-C data, so that 3D interactions can be functionally annotated. However, combining these integrations into a single figure is more intuitive and informative, and can better assist the interpretation of Hi-C contacts. We therefore develop two innovative plot functions, “HiC-integrationmap” and integration-track plot to better visualise Hi-C interactions and other data types.

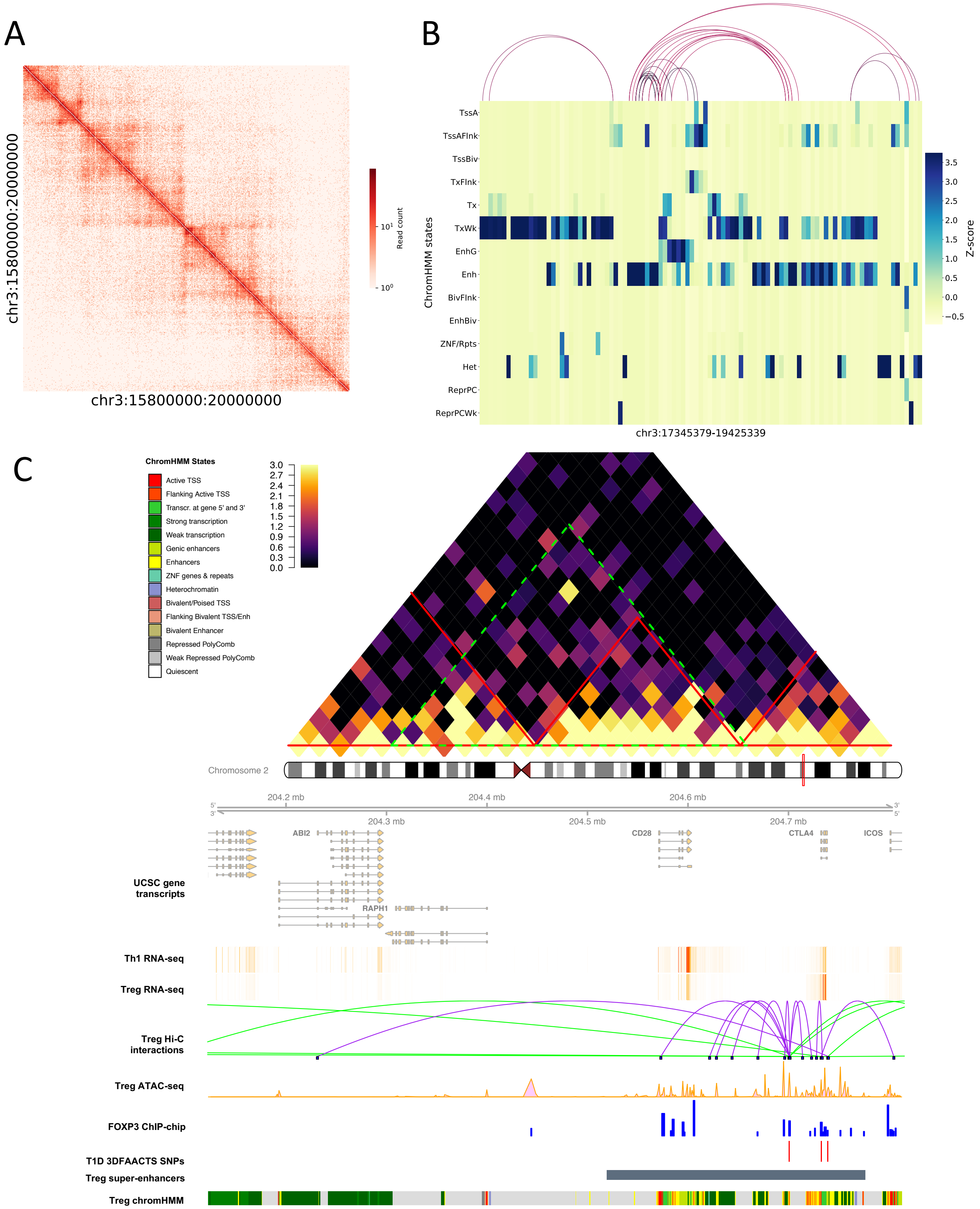

A
chr3:15800000:20000000
chr3:15800000:20000000
Read count
10^1
10^0
B
TssA
TssAFlnk
TssBiv
TxFlnk
Tx
TxWk
EnhG
Enh
BivFlnk
EnhBiv
ZNF/Rpts
Het
ReprPC
ReprPCWk
ChromHMM states
chr3:17345379-19425339
Z-score
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
C
ChromHMM States
Active TSS
Flanking Active TSS
Transcr. at gene 5' and 3'
Strong transcription
Weak transcription
Genic enhancers
Enhancers
ZNF genes & repeats
Heterochromatin
Bivalent/Poised TSS
Flanking Bivalent TSS/Enh
Bivalent Enhancer
Repressed PolyComb
Weak Repressed PolyComb
Quiescent
3.0
2.7
2.4
2.1
1.8
1.5
1.2
0.9
0.6
0.3
0.0
Chromosome 2
204.2 mb
204.4 mb
204.6 mb
204.3 mb
204.5 mb
204.7 mb
5'
3'
ABI2
RAPH1
CD28
CTLA4
ICOS
UCSC gene transcripts
Th1 RNA-seq
Treg RNA-seq
Treg Hi-C interactions
Treg ATAC-seq
FOXP3 ChIP-chip
T1D 3DFAACTS SNPs
Treg super-enhancers
Treg chromHMM

Figure 5: Visualisation methods of Hi-C data. A: using heatmap to visualise Hi-C interactions, the sequencing read pair count mapped to each interacting bin pair was used to determine the intensity of interactions, darker the color would indicate stronger interaction. B: An example of a HiC-integrationmap. The arc plot at the top displays the Hi-C interactions in this region, the heatmap underneath shows overlapping levels of each chromHMM state. C: An example of integration-tracks plot. The red triangles in the heatmap indicate Topologically Associated Domains (TADs) and the large green-dotted triangle indicates the boundary of the current plot. Tracks displayed below the chromosome 2 ideogram display integrating datasets along with various types of cell type-specific data including UCSC Gene Transcript information, T cell subsets (T-helper1 and Treg) expression data, Treg super-enhancer sets and 15-state ChromHMM track.

Tissue and cell type-specific chromHMM states data, which is an annotation of the non-coding region of the genome, are widely used in integration with Hi-C interactions to annotate potential functional interactions (Schmitt *et al.*, 2016; Greenwald *et al.*, 2018; Liu *et al.*, 2020). These chromHMM states data are predicted by a hidden markov model with five histone modification marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3) (Roadmap Epigenomics Consortium *et al.*, 2015). Therefore, a novel visualisation method called HiC-integrationmap was developed (https://github.com/ningbioinfostruggling/HiCvisualisation).

HiC-integrationmap is able to generate a heatmap with Hi-C interaction annotated above (Figure 5B) by defining a specific region and inputting the binned Hi-C interactions along with the appropriate chromHMM states data. In the heatmap, each column is the integration status of a bin, some of which are the anchors of the Hi-C interactions, and each cell is the z-score normalised overlapping bases of that bin with the chromHMM states. This heatmap is particularly useful in seeking potentially functional Hi-C interactions. For example, in a 2Mb plotting region (chr3:17-19Mb) (Figure 5B), we can observe dominant overlapping signals from the enhancer (Enh) and generic enhancer (EnhG) states, which indicate enhancers, along with Hi-C interactions linking to 2_TssAFlnk and 1_TssA overlapping areas, indicating promoters. Together they suggest that there are promoter-enhancer interactions that might govern the regulation of the gene in this region.

Besides chromHMM states, other epigenomics data are useful for Hi-C data integration. These may include data from sources such as ATAC-seq, RNA-seq, ChIP-seq of important transcription factors, single nucleotide polymorphism (SNP) and gene annotation. Taking advantages of published packages from bioconductor (Gentleman *et al.*, 2004), we combined R packages Gviz, which is useful for visualising genomic data as tracks, GenomicInteractions, which can process Hi-C data into track format, and coMET, which can generate color track for chromHMM states data visualisation, to develop an integration-tracks plot (Figure 5C). One of the most important advantages of the integration-tracks plot is that it allows visualisation of as much integration data as the user requires simultaneously.

An illustrative example shown in Figure 5C displays overlapping regulatory T cell (Treg) specific data and three type 1 diabetes-specific SNPs in the plotting area with overlapped ATAC-seq signals and transcription factor FOXP3-binding sites.

All of these datasets overlap with gene CTLA4, indicating that these SNPs might be important for the regulation of CTLA4, and further that this regulation tends to be Treg-specific. Additionally, we also developed an R function HiCheatmap (https://github.com/ningbioinfostruggling/HiCvisualisation) to generate the classic heatmap as triangle shape (Figure 5C) to visualise the interaction intensity in the plotting region, with optional indication of predicted topologically-associated domains (TADs).

Discussion

In this chapter, we reviewed and summarised four standardised steps commonly used to process Hi-C data in published studies, including aligning sequencing reads to reference genome, filtering uninformative read pairs, generating interaction matrix and matrix normalisation. However, we also demonstrated that there are three aspects of the process that are able to be further optimized.

Assessment of Hi-C data quality is an important part of reproductive research, however, no tools currently exist to extract and summarise quality statistics from standard Hi-C analysis pipelines. To overcome this limitation we developed a computational pipeline called HiC-QC that is able to summarise QC statistics and output these metrics to file in either csv or Excel format depending on the user requirement. However, there are still limitations in both HiC-QC and the field of quality control of Hi-C data. Currently, HiC-QC can only be used with the results of HiC-Pro, one of most popular pipelines. The ability to include additional input types will be addressed in future releases, specifically when using pipelines, such as Juicer (Durand et al., 2016) and FAN-C (Kruse, Hug and Vaquerizas, 2020), that rely on the BWA alignment. This will be addressed in a future release, where we will consider output structures of other commonly used pipelines, especially for pipelines using BWA aligner, which we found is a more appropriate option when analysing DNase Hi-C data, such as Juicer (Durand *et al.*, 2016) and FAN-C (Kruse, Hug and Vaquerizas, 2020). Additionally, with more and more Hi-C derived protocols being developed, such as HiChIP (Mumbach *et al.*, 2016), BL-Hi-C (Liang *et al.*, 2017), Ocean-C (Li *et al.*, 2018), DLO Hi-C (Lin *et al.*, 2018), tagHi-C (Zhang *et al.*, 2020) and scHi-C (Nagano, Wingett and Fraser, 2017), data generated by different protocols may require protocol-specific statistics, such as capture efficiency and on-target rate for capture Hi-C and HiChIP, to better evaluate the quality of the library, and the recommended ranges for specific statistics may need adjustment for different datasets.

Further to the development of the new Hi-C-QC tool, we compared Bowtie2 and BWA, two of the most popular choices for aligning Hi-C data, by investigating their performance in treatment of low mapping rate Hi-C data, normal *in situ* Hi-C data, capture Hi-C data and DNase Hi-C data. We found that when treating *in situ* Hi-C and capture Hi-C data, BWA and Bowtie2 identify a similar number of Hi-C interactions with over 80% of the interactions overlapping (Figure 2). BWA outperformed Bowtie2 when aligning DNase Hi-C data and low mapping quality data. Currently, the alignment of Hi-C data to reference genomes such as human hg38 and mouse mm10 genome has been popularly utilised, however graph-based reference genomes have become more and more popular because they allow sequence read mapping to exact haplotypes, and there are recent studies showing that mapping to graph-based genomes can improve accuracy compared to linear genomes (Garrison *et al.*, 2018; Rakocevic *et al.*, 2019). Therefore, in the future, developing a graph-based alinger for Hi-C data may benefit the society of researchers investigating 3D genome structure and allow further understanding of the mechanism of how 3D interactions govern gene regulation.

Last but not least, current existing visualisation methods of Hi-C data often neglect data integration, which can assist the functional interpretation of chromatin interactions. To improve visualisation and interpretation of Hi-C interaction, we therefore developed HiC-integrationmap and integration-track plot, which are able to assist in the visualisation of Hi-C interactions as well as with the integrations with other epigenomics data. However, a limitation of both of these visualisation tools is the need to input only the specific region of the genome to be inspected. This highlights a classic dilemma in the visualising of Hi-C data where we either visualise the whole genome but it is overwhelming to add integration visualisations, or we only visualise a specific region, but it requires prior information regarding the region of interest. One potential solution would be to develop an interactive browser for Hi-C data, similar to the UCSC browser (Kent et al., 2002) or the WashU Epigenome browser (Li et al., 2019). Such interactive browser version of HiC-integrationmap and integration-track plot can be developed by R shiny app or python Dash framework in the future. Additionally, interactive visualisation of large amounts of data integration requires a large investment in computational resources. This limitation could be addressed by the utilisation of a cloud computing platform such as those hosted by Amazon Web Services, Google cloud platform and Microsoft Azure.

References

Akdemir, K. C. and Chin, L. (2015) ‘HiCPlotter integrates genomic data with interaction matrices’, *Genome biology*, 16, p. 198.

Andrews, S. and Others (2010) ‘FastQC: a quality control tool for high throughput sequence data’. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Ay, F., Bailey, T. L. and Noble, W. (2014) ‘Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts’, *Genome research*, 24(6), pp. 999–1011.

Barisic, D. *et al.* (2019) ‘Mammalian ISWI and SWI/SNF selectively mediate binding of distinct transcription factors’, *Nature*, 569(7754), pp. 136–140.

Belaghzal, H., Dekker, J. and Gibcus, J. H. (2017) ‘Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation’, *Methods* , 123, pp. 56–65.

Bertero, A. *et al.* (2019) ‘Dynamics of genome reorganization during human cardiogenesis reveal an RBM20-dependent splicing factory’, *Nature communications*, 10(1), p. 1538.

Dekker, J. *et al.* (2017) ‘The 4D nucleome project’, *Nature*, 549(7671), pp. 219–226.

Deng, X. *et al.* (2015) ‘Bipartite structure of the inactive mouse X chromosome’, *Genome biology*, 16, p. 152.

Dixon, J. R. *et al.* (2012) ‘Topological domains in mammalian genomes identified by analysis of chromatin interactions’, *Nature*, 485(7398), pp. 376–380.

Dixon, J. R. *et al.* (2015) ‘Chromatin architecture reorganization during stem cell differentiation’, *Nature*, 518(7539), pp. 331–336.

Durand, N. C. *et al.* (2016) ‘Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments’, *Cell systems*, 3(1), pp. 95–98.

Ewels, P. *et al.* (2016) ‘MultiQC: summarize analysis results for multiple tools and samples in a single report’, *Bioinformatics* , 32(19), pp. 3047–3048.

Garrison, E. *et al.* (2018) ‘Variation graph toolkit improves read mapping by representing genetic variation in the reference’, *Nature biotechnology*, 36(9), pp. 875–879.

Gentleman, R. C. *et al.* (2004) ‘Bioconductor: open software development for computational biology and bioinformatics’, *Genome biology*, 5(10), p. R80.

Greenwald, W. W. *et al.* (2018) ‘Pancreatic islet chromatin accessibility and conformation defines distal enhancer networks of type 2 diabetes risk’, *BioRxiv*, p. 299388.

Hu, M. *et al.* (2012) ‘HiCNorm: removing biases in Hi-C data via Poisson regression’, *Bioinformatics* , 28(23), pp. 3131–3133.

Hwang, Y.-C. *et al.* (2015) ‘HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements’, *Bioinformatics* , 31(8), pp. 1290–1292.

Imakaev, M. *et al.* (2012) ‘Iterative correction of Hi-C data reveals hallmarks of chromosome organization’, *Nature methods*, 9(10), pp. 999–1003.

Javierre, B. M. *et al.* (2016) ‘Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters’, *Cell*, 167(5), pp. 1369–1384.e19.

Jin, F. *et al.* (2013) ‘A high-resolution map of the three-dimensional chromatin interactome in human cells’, *Nature*, 503(7475), pp. 290–294.

Jung, I. *et al.* (2019) ‘A compendium of promoter-centered long-range chromatin interactions in the human genome’, *Nature genetics*, 51(10), pp. 1442–1449.

Kent, W. J. *et al.* (2002) ‘The human genome browser at UCSC’, *Genome research*, 12(6), pp. 996–1006.

Klocko, A. D. *et al.* (2016) ‘Normal chromosome conformation depends on subtelomeric facultative heterochromatin in Neurospora crassa’, *Proceedings of the National Academy of Sciences of the United States of America*, 113(52), pp. 15048–15053.

Knight, P. A. and Ruiz, D. (2013) ‘A fast algorithm for matrix balancing’, *IMA Journal of Numerical Analysis*, 33(3), pp. 1029–1047.

Kruse, K., Hug, C. B. and Vaquerizas, J. M. (2020) ‘FAN-C: A Feature-rich Framework for the Analysis and Visualisation of C data’, *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/2020.02.03.932517v1.abstract.

Lajoie, B. R., Dekker, J. and Kaplan, N. (2015) ‘The Hitchhiker’s guide to Hi-C analysis: practical guidelines’, *Methods* , 72, pp. 65–75.

Liang, Z. *et al.* (2017) ‘BL-Hi-C is an efficient and sensitive approach for capturing structural and regulatory chromatin interactions’, *Nature communications*, 8(1), p. 1622.

Li, D. *et al.* (2019) ‘WashU Epigenome Browser update 2019’, *Nucleic acids research*, 47(W1), pp. W158–W165.

Lieberman-Aiden, E. *et al.* (2009) 'Comprehensive mapping of long-range interactions reveals folding principles of the human genome', *Science*, 326(5950), pp. 289–293.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics* , 25(16), pp. 2078–2079.

Li, H. (2013) 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM', *arXiv [q-bio.GN]*. Available at: http://arxiv.org/abs/1303.3997.

Lin, D. *et al.* (2018) 'Digestion-ligation-only Hi-C is an efficient and cost-effective method for chromosome conformation capture', *Nature genetics*, 50(5), pp. 754–763.

Li, T. *et al.* (2018) 'OCEAN-C: mapping hubs of open chromatin interactions across the genome reveals gene regulatory networks', *Genome biology*, 19(1), p. 54.

Liu, N. *et al.* (2020) '3DFAACTS-SNP: Using regulatory T cell-specific epigenomics data to uncover candidate mechanisms of Type-1 Diabetes (T1D) risk', *Cold Spring Harbor Laboratory*. doi: 10.1101/2020.09.04.279554.

Lucic, B. *et al.* (2019) 'Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration', *Nature communications*, 10(1), p. 4059.

Lun, A. T. L. and Smyth, G. K. (2015) 'diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data', *BMC bioinformatics*, 16, p. 258.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), pp. 10–12.

Ma, W. *et al.* (2018) 'Using DNase Hi-C techniques to map global and local three-dimensional genome architecture at high resolution', *Methods* , 142, pp. 59–73.

Mifsud, B. *et al.* (2015) 'Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C', *Nature genetics*, 47(6), pp. 598–606.

Mifsud, B. *et al.* (2017) 'GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data', *PloS one*, 12(4), p. e0174744.

Mumbach, M. R. *et al.* (2016) 'HiChIP: efficient and sensitive analysis of protein-directed genome architecture', *Nature methods*, 13(11), pp. 919–922.

Nagano, T., Wingett, S. W. and Fraser, P. (2017) 'Capturing Three-Dimensional Genome Organization in Individual Cells by Single-Cell Hi-C', *Methods in molecular biology* , 1654, pp. 79–97.

Oksuz, B. A. *et al.* (no date) 'Systematic evaluation of chromosome conformation capture assays'. doi: 10.1101/2020.12.26.424448.

Pal, K., Forcato, M. and Ferrari, F. (2019) 'Hi-C analysis: from data generation to integration', *Biophysical reviews*, 11(1), pp. 67–78.

Patel, R. K. and Jain, M. (2012) 'NGS QC Toolkit: a toolkit for quality control of next generation sequencing data', *PloS one*, 7(2), p. e30619.

Rakocevic, G. *et al.* (2019) 'Fast and accurate genomic analyses using genome graphs', *Nature genetics*, 51(2), pp. 354–362.

Ramani, V. *et al.* (2016) 'Mapping 3D genome architecture through in situ DNase Hi-C', *Nature protocols*, 11(11), pp. 2104–2121.

Rao, S. S. P. *et al.* (2014) 'A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping', *Cell*, 159(7), pp. 1665–1680.

Rieber, N. *et al.* (2013) 'Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies', *PloS one*, 8(6), p. e66621.

Roadmap Epigenomics Consortium *et al.* (2015) 'Integrative analysis of 111 reference human epigenomes', *Nature*, 518(7539), pp. 317–330.

Rodrigues, P. *et al.* (2018) 'NF-κB-Dependent Lymphoid Enhancer Co-option Promotes Renal Carcinoma Metastasis', *Cancer discovery*, 8(7), pp. 850–865.

Schmitt, A. D. *et al.* (2016) 'A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome', *Cell reports*, 17(8), pp. 2042–2059.

Schubert, M., Lindgreen, S. and Orlando, L. (2016) 'AdapterRemoval v2: rapid adapter trimming, identification, and read merging', *BMC research notes*, 9, p. 88.

Servant, N. *et al.* (2015) 'HiC-Pro: an optimized and flexible pipeline for Hi-C data processing', *Genome biology*, 16, p. 259.

Sinkhorn, R. and Knopp, P. (1967) 'Concerning nonnegative matrices and doubly stochastic matrices', *Pacific Journal of Mathematics*, 21(2), pp. 343–348.

Taberlay, P. C. *et al.* (2016) 'Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations', *Genome research*, 26(6), pp. 719–731.

Vidal, E. *et al.* (2018) 'OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes', *Nucleic acids research*, 46(8). doi: 10.1093/nar/gky064.

Vieux-Rochas, M. *et al.* (2015) 'Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain', *Proceedings of the National Academy of Sciences of the United States of America*, 112(15), pp. 4672–4677.

Ward, C. M., To, T.-H. and Pederson, S. M. (2020) 'ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files', *Bioinformatics* , 36(8), pp. 2587–2588.

Wingett, S. *et al.* (2015) 'HiCUP: pipeline for mapping and processing Hi-C data', *F1000Research*, 4, p. 1310.

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

Yang, L. *et al.* (2020) ‘3D Genome Analysis Identifies Enhancer Hijacking Mechanism for High-Risk Factors in Human T-Lineage Acute Lymphoblastic Leukemia’, *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/2020.03.11.988279v1.abstract.

Zhang, C. *et al.* (2020) ‘tagHi-C Reveals 3D Chromatin Architecture Dynamics during Mouse Hematopoiesis’, *Cell reports*, 32(13), p. 108206

Chapter 3

3DFAACTS-SNP: Using regulatory T cell-specific epigenomics data to uncover candidate mechanisms of Type-1 Diabetes (T1D) risk

3DFAACTS-SNP: Using regulatory T cell-specific epigenomics data to uncover candidate mechanisms of Type-1 Diabetes (T1D) risk

Ning Liu[1,2,3,+], Timothy Sadlon[2,4,+], Ying Ying Wong[2,4], Stephen Pederson[3], James Breen[1,2,3,*,#] & Simon C Barry[2,4,#]

[1] South Australian Health & Medical Research Institute, Adelaide, Australia

[2] Robinson Research Institute, University of Adelaide, Adelaide, Australia

[3] Bioinformatics Hub, School of Biological Sciences, University of Adelaide, Adelaide, Australia

[4] Women's & Children's Health Network, Adelaide, Australia

[+] These authors contributed equally

[#] These authors jointly directed this work

* Corresponding authors: James Breen (jimmy.breen@sahmri.com)

Abstract

Background

Genome-wide association studies (GWAS) have enabled the discovery of single nucleotide polymorphisms (SNPs) that are significantly associated with many autoimmune diseases including type 1 diabetes (T1D). However, many of the identified variants lie in non-coding regions, limiting the identification of

mechanisms that contribute to autoimmune disease progression. To address this problem, we developed a variant filtering workflow called 3DFAACTS-SNP to link genetic variants that are associated with T1D to the loss of immune tolerance in regulatory T cells (Treg).

Results

Using 3DFAACTS-SNP we identified 36 SNPs with plausible Treg-specific mechanisms of action contributing to T1D from 1,228 T1D fine-mapped variants, identifying 119 novel interacting regions resulting in the identification of 51 candidate target genes. We further demonstrated the utility of the workflow by applying it to three other meta-analysed SNP autoimmune datasets, identifying 17 Treg-centric candidate variants and 35 interacting genes. Finally, we demonstrate the broad utility of 3DFAACTS-SNP for functional annotation of all known common (>10% allele frequency) variants from the Genome Aggregation Database (gnomAD). We identified 7,900 candidate variants and 3,245 candidate target genes, generating a list of potential sites for future T1D or autoimmune research.

Conclusions

We demonstrate that it is possible to further prioritise variants that contribute to T1D based on regulatory function and illustrate the power of using cell type specific multi-omics datasets to determine disease mechanisms. Our workflow can be customised to any cell type for which the individual datasets for functional annotation have been generated, giving broad applicability and utility.

Keywords

Type 1 diabetes; Autoimmune disease; Regulatory T cells; Hi-C; Transcription factor binding; Functional annotation.

Background

Autoimmune diseases are chronic inflammatory disorders caused by a breakdown of immunological tolerance to self-antigens, which results in an imbalance between multiple immune cells, including conventional T cells (Tconvs) and regulatory T cells (Tregs) (1). The imbalance of immune cell function can lead to the destruction of host tissues, such as is observed in multiple autoimmune diseases, including rheumatoid arthritis (RA) (joint tissues), multiple sclerosis (MS) (myelinated nerves) and inflammatory bowel disease (IBD) (intestine /colon). In the case of Type 1 Diabetes (T1D), a reduction of Treg cell function contributes to unrestrained immune destruction of the insulin-generating pancreatic beta cells (2).

Regulatory T cell function is mediated by expression of the Foxhead Box Protein 3 (FOXP3) transcription factor (TF) as evidenced by severe autoimmune diseases observed in FOXP3-deficient scurfy mice (3) and IPEX in humans (4–6). RNA sequencing and chromatin immunoprecipitation (ChIP) studies have uncovered an extensive FOXP3-dependent molecular program involved in Treg cell development and stability (7,8), and functional fitness of Treg is dependent on stable robust expression of FOXP3, such that reduced FOXP3 expression is linked to reduced Treg function. For example, in a small T1D cohort study, we have shown that there is a decrease in FOXP3 expression in the Treg of children over the first 9 months post diagnosis (9). However, since FOXP3 itself is not mutated in autoimmune diseases other than IPEX, the loss of FOXP3 levels and functional fitness is likely caused by perturbation of the Treg gene regulatory network. Hence, by decoding the regulatory network of FOXP3, and mapping the genetic risk to the key functional genes it impacts, we will gain a better understanding of how autoimmune diseases like T1D could be countered.

T1D occurs spontaneously in approximately 80% of individuals, however predisposition to the disease has a strong pattern of inheritance (10). Genome-Wide Association Studies (GWAS) have identified over 50 loci that are strongly associated with T1D, based on the genotyping of a total of 9934 cases and 16956 controls from multiple cohorts and resources (11). In addition, fine-mapping of immune-disease associated loci represented on the Immunochip Array (12) followed by a Bayesian approach identified 44 significant T1D-associated Loci and over 1,000 credible SNPs (13). While alterations in either the effector or regulatory arms of the immune system can result in loss of tolerance and autoimmune disease, we have used a Treg centric view of loss of tolerance. This is based on the observation that defects in Treg function have been reported in autoimmune diseases including T1D and MS (14,15) and that experimental deletion of FOXP3 or reduced Treg function results in autoimmune disease in many model systems (16,17).

Although GWAS have revealed significant associations between genetic variants and T1D, the vast majority of the sampled single nucleotide polymorphisms (SNPs) are located in non-coding regions that do not alter the amino acid sequence in a protein, making it difficult to assign direct biological functions to variants (18–20). Non-coding variants can be linked to direct changes in gene expression by identifying expression quantitative trait loci (eQTL) that aim to associate allelic changes to a cis (within 1Mbp of the associated gene) and trans (>1Mbp) change in gene expression (21,22). This additional direct gene expression association however still fails to identify direct mechanisms by which a specific genetic variant can change gene expression. In addition, usage of eQTLs to establish direct changes from GWAS variants is somewhat limited to local, or cis-eQTLs (23,24), whereas mounting evidence shows that long-range regulatory connections, driven by three-dimensional chromatin interactions (25,26), can mediate these changes in expression.

With the increasing affordability and availability of high-throughput sequencing techniques and various epigenomics sequencing data protocols, the impact of genome organization and accessibility can now be added to the functional annotation of genetic risk. Chromatin immunoprecipitation sequencing (ChIP-seq) allows us to identify the binding sites of a transcription factor; assay for transposase-accessible chromatin sequencing (ATAC-seq) data offers the ability to identify highly accessible regions of the genome; and high resolution chromosome conformation capture sequencing (Hi-C) data can facilitate the investigation of the three-dimensional structure of the genome. Since it is believed that the mechanisms by which non-coding SNPs contribute to diseases are mostly via changes to the function of regulatory elements (20), we believe that combining multiple genomics and epigenetics sequencing data can further reveal the relationship between GWAS SNPs and disease pathways. Our hypothesis is that the genetic variation that specifically alters Treg function will reside in open chromatin in Treg cells that is bound by FOXP3 and the genes controlled by these by regulatory regions can be identified by chromosome conformation capture approaches. Therefore, in this paper, we describe a filtering workflow using multiple sequencing data from human Tregs, aiming to identify plausible immunomodulatory mechanisms and potentially find previously unknown connections between causative variant SNPs significantly associated with T1D and the genes they impact.

Results

Post-GWAS filtering using Treg-specific epigenomic datasets prioritises functionally relevant genetic variants contributing to T1D

As T1D is partly a consequence of Treg dysfunction, we infer that variants contained within active regulatory regions of Treg cells are likely to contribute to disease progression by impacting Treg function. A view supported by the finding that T1D associated SNPs are enriched at Treg-specific regulatory regions (27). Therefore, starting with published T1D GWAS variant information, we designed a filtering workflow (Figure 1) using multiple human Treg-specific epigenomic data to identify perturbations within defined “regulatory T cell active regions”.

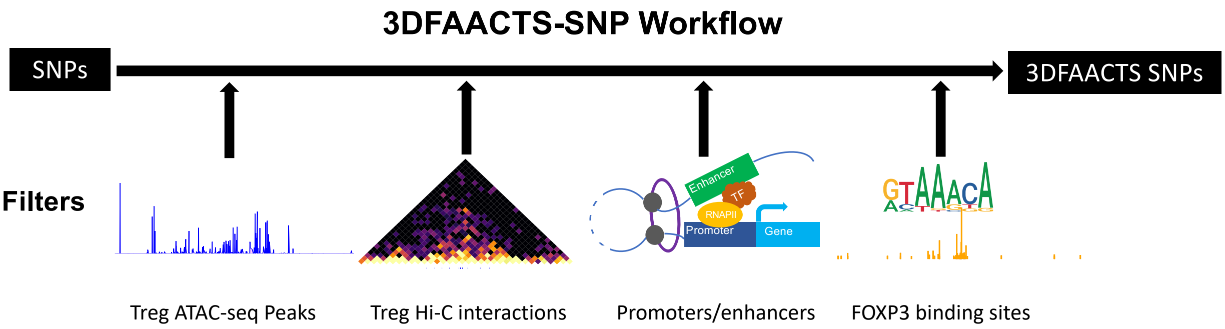


Figure 1: Diagram of the individual components of the Treg-specific 3DFAACTS-SNP filtering workflow for identifying variants that are potentially causative to Type 1 Diabetes (T1D). GWAS or fine-mapped variants (on the left) are intersected with different filtering elements, including Treg ATAC-seq peaks, interactions from Treg Hi-C, promoters or enhancers and previously identified FOXP3 binding regions in Treg cells (28), resulting in filtered variants we termed 3DFAACTS SNPs.

In order to obtain highly accessible chromatin regions in Treg, we performed Transposase-Accessible Chromatin using sequencing (ATAC-seq) on resting and stimulated Treg cells from three donors and sequenced to an average of 37.1 million reads (± 4 million) per sample. From the ATAC-seq data, we identified 525,647 ATAC-seq peaks on average (Additional file1: Table S1). These ATAC-seq peaks were then merged into 683,954 non-redundant peaks and used to screen for variants located in accessible regions in regulatory T cells as the first filtering step of the 3DFAACTS-SNP pipeline (Figure 1).

Numerous studies have shown that three dimensional (3D) interactions play important roles in gene regulation, mediated by DNA looping bringing enhancers and promoters together at transcriptional hubs (29–31). As a result, distant loci which physically interact with disease associated regulatory regions can be potentially impacted by these regions. To identify 3D interacting regions in Treg cells, we generated and sequenced Treg *in situ* Hi-C libraries. Two technical replicates of human Treg Hi-C libraries were sequenced to an average depth of 3 million reads, and after processing using HiC-Pro (32) and quality control by HiC-QC, generated 459,244 and 1,441,362 Hi-C valid interactions respectively (Additional file1: Table S2). We extended these interactions to form 2000bp (+/-1000bp upstream and downstream) windows at both ends of each interaction. We then collapsed interactions by merging interactions with overlapping anchors to generate non-redundant interaction pairs which represent Hi-C interactions in Tregs. These non-redundant interactions were then integrated with the variant associated ATAC-seq peaks identified above to identify accessible interacting regions.

To assign potential function to identified variant associated ATAC-seq peaks and Hi-C interacting regions we next determined the overlap of these regions with enhancer and promoter annotations (Figure 1). This included 113,369 enhancers (mean size of 698bp) identified by the Functional Annotation of the Mammalian Genome (FANTOM5) project (33) and promoter regions (n = 73,171) associated with GRCh37/hg19 UCSC known transcripts. Promoters were defined by extending upstream 2 kb of transcription start sites (TSS). In addition, we extended the list of regulatory regions using the 15 state chromHMM model for CD4+ CD25+ CD127- Primary Treg cells from the Roadmap Epigenomics Project (34). We defined chromHMM states *EnhG*, *Enh* and *EnhBiv* as enhancers and *TssA*, *TssAFlnk*, *TssBiv* and *BivFlnk* as promoters. FANTOM5 enhancers and defined promoters and chromHMM enhancers/promoters states were then merged respectively to represent all possible genetic regulatory elements, covering 7.49 % of the genome (Additional file2: Table S3).

The transcription factor FOXP3 is critical for Treg function and orchestrating immunological tolerance, and stable high FOXP3 expression levels are observed specifically in Tregs (3,28,35). Therefore, by intersecting filtered SNPs with significant human FOXP3-binding signals, we can largely constrain SNPs within regulatory regions to FOXP3 controlled Treg-specific gene networks (Figure 1). We used 8,304 (mean size = 1317bp) FOXP3 ChIP-chip peaks from our previous study (28) to specify FOXP3 binding in humanTreg cells. Of interest, by searching the Gene & Autoimmune Disease Association Database (GAAD) (36), we obtained 245 annotated genes that are associated with T1D, and found a significant enrichment of FOXP3 binding sites in T1D-associated genes (Fisher exact test: P-value = 4.519e-09), suggesting a strong association between T1D risk and FOXP3 controlled Treg function. Taken together, FOXP3 binding, physical interaction, regulatory element and open chromatin regions offer a large subset of regions to use for GWAS variant prioritisation and functional annotation experiments.

Linking fine-mapped T1D-associated variants to their targets via chromatin interactions

Genetic studies have identified over 50 candidate gene regions that contain potentially causative SNPs that impact T1D (11). Recently, a study of T1D-associated variants using Immunochip, a custom-made SNP array containing immune-related genetic variants from the 1000 genomes project (12,37), and Bayesian fine-mapping identified 1,228 putative causal variants associated with T1D (13). We used our workflow to further prioritise variants from this fine-mapped set to investigate potentially causative SNPs that contribute to T1D via affecting promoter/enhancer interaction in human Treg cells.

Table 1: T1D 3DFAACTS SNPs identified using the 3DFAACTS-SNP filtering workflow from T1D fine-mapping SNPs (13)**.** The nearest locus indicates the closest gene to the variants in linear distance, while 3D interacting genes are genes contact with the variants via Treg Hi-C interactions. Overlapped regulatory elements of each 3DFAACTS SNPs are displayed, including chromatin states from a 15-states model (34) and expressed enhancers from FANTOM5 (38). Detailed SNP and interaction information is contained in Supplementary Information (Additional file 3: Table S4).

Chromosome	Position	SNP id	Nearest Locus (linear distance)	Located within regulatory regions		Interacting Genes (3D)
				Treg ChromHMM	FANTOM5 expressed enhancers	
chr2	204700689	rs12990970	CTLA4	TssAFlnk		**TLK1**,**NBEAL1**,**CD28**
	204732714	rs231775		TssAFlnk		**KIAA2012**,**ICOS**

	204738919	rs3087243		EnhG		**ABI2**,**IQCA1**
chr3	46327588	rs11718385	CCR3	Enh		
	46391390	rs6441972	CCR2	TssAFlnk		
	46401032	rs3138042		Enh		**MLH1**,**LRRFIP2**,CCR2
	46411661	rs2856758	CCR5	Enh		**CCR3**,**TIPARP**,**KLHL24**
	46412259	rs1799988		TssAFlnk		**CCR3**,**TIPARP**,**KLHL24**
chr5	35852311	rs6890853	IL7R	TssAFlnk		**SPEF2**
chr6	90948476	rs62408222	BACH2	Enh		
	90983850	rs905671		Enh	✓	**ZNF292**,**ANKRD6**,**LYRM2**
	90984035	rs943689		Enh	✓	**ANKRD6**,**LYRM2**
	90995980	rs614120		TssAFlnk	✓	BACH2,**AFG1L**
chr7	50462418	rs10216316	IKZF1	EnhG		IKZF1,**GRB10**
	50462498	rs10215297		EnhG		IKZF1,**GRB10**
	50465206	rs55981617		EnhG		**DPY19L2P3**,IKZF1,**DDC**,**CNOT4**
	50465654	rs12670555		EnhG		**DPY19L2P3**,IKZF1,**DDC**,**CNOT4**
chr10	6088743	rs12722508	IL2RA	TssAFlnk		IL2RA,**PFKFB3**,**HECTD2-AS1**
	6094697	rs61839660		TssAFlnk		**SFMBT2**
	6096667	rs12722496			✓	IL2RA,**RBM17**,**PFKFB3**,**LINC02649**, **PPA1**,**BORCS7-ASMT**
	6107534	rs11597367		Enh		**IL15RA**,IL2RA,**SFMBT2**
chr12	9910720	rs3176793	CD69	TssA		CD69
	9912182	rs2160086		TssA		**CLEC2D**,CD69,**CLEC2A**
	9912730	rs3176789		TssA		**CLEC2D**,CD69,**CLEC2A**
	9916640	rs3136559		Enh		CD69,**YBX3**
	9925758	rs1029992		Enh		**CHD4**,**BORCS5**
	9926064	rs1029991		Enh		**CHD4**
	9926397	rs1029990		Enh	✓	**CHD4**,**CLEC2D**

	9926624	rs10844749		Enh		**CHD4**,**CLEC2D**
	9926784	rs1540356		Enh		**CHD4**,**CLEC2D**
chr15	38903672	rs16967112	RASGRP1	Enh	✓	RASGRP1,**CHP1**, **DNAAF4-CCPG1**,**ZNF592**
	38903884	rs56249992		Enh		RASGRP1,**CHP1**,**DNAAF4-CCPG1**
chr16	11188949	rs71136618	CLEC16A	Enh		**RMI2**
chr17	38755665	rs11656173	SMARCE1	Enh	✓	**RARA**,**TOP2A**
chr18	12838767	rs17657058	PTPN2	Enh		**SPIRE1**,**PIGN**
chr22	30581722	rs5753037	HORMAD2	Enh		

*Note: Genes in bold indicate novel 3D interacting genes of the identified SNPs.

From the 1,228 fine-mapped T1D-associated SNPs, we identified 36 variants that meet our filtering criteria as described above, in this study we will refer to them as T1D 3DFAACTS SNPs. These variants are located at 14 different chromosomal loci and distally interact with a further 80 regions in Tregs (Table 1 & Additional file 3: Table S4). The majority of variants (71.4%, 25 out of 35 SNPs) were located in enhancer regions rather than promoters while one variant, rs614120 is located in both the *TssAFlnk* chromHMM state and T cell-specific enhancers from FANTOM5. Given that a *TssAFlnk* state can either indicate a promoter or enhancer (39), combining with the identified FANTOM enhancer information we believe that rs614120 is more likely to be located within an enhancer region. This observed variant enrichment over enhancer states may be caused by an uneven number of promoters and enhancers used in the filtering scheme, where the accumulated bases of enhancers (77,217,165 bp) is significantly larger than promoters (71,634,647 bp). Another plausible explanation of this bias is that risk variants of T1D are more likely to be in enhancer regions and influence the transcriptional output via affect the function of enhancers. This is consistent with the summary in previous review (173) that the number of found genetic risk variants that affect enhancer function is estimated to be much larger than the ones that impact promoter function. Of the 14 loci identified, 8 contained more than two plausible variants across the loci. For example, variants located near the CD69 gene on chromosome 12 had the highest number of filtered variants, with 9 variants located in regulatory regions around the gene. In order to annotate the filtered variants to nearby genes, we took two approaches: annotated genes that were located in proximity to the SNPs using linear, chromosomal distances, and genes identified by their interaction with variant-containing regulatory regions via Treg Hi-C interactions (Table 1). Genes proximal to the identified 36 T1D variants include CTLA4, CCR5, IL7R, BACH2, IKZF1, IL2RA, CD69, RASGRP1, CCR3, CCR2, CLEC16A, HORMAD2 and PTPN2. These genes have previously been associated with T1D (13) and in addition other autoimmune disorders such as Multiple Sclerosis (MS), Rheumatoid Arthritis (RA), Crohn’s Disease (CD) and Inflammatory Bowel Disease (IBD) (40–44). Additionally, we annotated the filtered variants using eQTL data across all tissues from the Genotype-Tissue Expression (GTEx) project (45) and immune cells using the DICE database (46). We found that 12 filtered SNPs are annotated as the eQTL to their nearest loci (Additional file 3: Table S4) while 4 SNPs, rs11718385 (CCR3), rs62408222, rs905671 and rs943689 (BACH2) were identified as eQTL to their nearest gene (nondirected) in Tregs (46). These data confirmed the ability of 3DFAACTS-SNP to identify potential disease associated regulatory region-target gene networks in a cell type specific manner.

In addition to the annotation of the 36 T1D SNPs to 14 genes in closest linear proximity, 3DFAACTS-SNP identified 119 interacting regions and a further 51 genes that interact with the variant containing regulatory regions via Treg Hi-C (Table 1 & Additional file 3: Table S4). We next used the 15 states regulatory model for CD4+ CD25+ CD127- Treg primary cells from the Roadmap Epigenomics Project (34) to annotate interacting regions. These regions most frequently overlapped active chromatin states associated with transcription and gene regulation including states associated with weak transcription (*5_TxWk*) in 30% of identified regions, enhancers (*7_Enh*) in 29%, flanking active TSS (*2_TssAFlnk*) in 21% and 13% of regions located in active TSS state (*1_TssA*) (Additional file 3: Table S4). Two genes, *DPY19L2P3* and *DDC* were then dropped from further analysis as they did not overlap active states in a Treg. Additionally, searches of the GAAD (36) indicated that 45 % (22/49) of the 3D interacting genes have been previously associated with autoimmune diseases including Rheumatoid arthritis, Multiple sclerosis, Inflammatory bowel disease and T1D (Additional file 3: Table S4). Of these 22 interacting genes, 6 have been shown to be significantly associated with T1D, including BACH2, CD28, CD69, ICOS, IL2RA and RASGRP1 (Additional file 3: Table S4). Overall, by overlapping with chromHMM states, we found 49 genes and 80 interacting regions that are active in Tregs that are in close proximity to regulatory regions carrying TD-associated variants.

Taken together, our analysis identified 31 new T1D candidate genes that may be disrupted in Treg, and a further 18 genes that have been previously associated with T1D (13). Furthermore, 61% of these interacting regions and 13 genes overlap with induced Treg super-enhancers (SEs; http://www.licpathway.net/sedb/), consistent with these regions containing important Treg functional elements. When looking at the mean normalised expression (FPKM > 1) of genes in Treg samples in Gao et al 2019 (47), 78% of interacting genes (Additional file 3: Table S4) are expressed in Tregs, all of which were enriched for T cell specific gene ontologies (Additional file 1: Figure S1). These data indicate that distal interacting regions contain regulatory regions and genes important for Treg function and are consistent with a model in which the variant containing regulatory regions may contribute to T1D by disrupting the regulation of these distal interacting genes.

The topological neighbourhood surrounding filtered T1D variants

We next investigated the topological neighbourhood, i.e. the presence of topologically-associated or frequently interacting domains, in which regulatory regions harbouring the filtered T1D variants reside. By establishing putative boundaries of each 3D structural domain, we are then able to characterise the coordination of contacts within a loci and how they act to control gene expression. We called topologically-associated domains (TADs) using Treg Hi-C data (Additional file 4: Table S5) used in the workflow described above and integrated with publicly available super-enhancer, chromHMM data of T cell lineages and Treg expression data (48). All data was overlapped across each locus and displayed in supplementary figures 2-13.

TADs are called based on the frequency of interactions within a region (49), with physical interactions between two loci generally decaying with increasing linear distance on the chromosome (50). Genes in the closest proximity to our filtered T1D variants (Table 1), were unsurprisingly found within the same TAD. Interestingly however, we found that interacting regions and genes identified by Hi-C were only co-located within the same TAD in ~56.5% of cases (i.e. intra-TAD interactions), with 42.6% of interactions occurring between different TADs (inter-TAD; Additional file 4: Table S5). Indeed, the linear distance between filtered variants and their 3D interacting genes (~12.5Mb) were on average ~2.3Mb further away compared to the average distance of intrachromosomal interactions found in the entire Treg Hi-C dataset (~10.2Mb), indicating that Treg-active, FOXP3-bound regions impact genes across much greater linear distances than regular connections.

A high degree of chromatin interactions between genes and enhancer regions was detected within the filtered variant containing TADs, with these interactions both confirming previously identified SNP-target combinations and indicating potential new targets for investigation. For example, 3DFAACTS-SNP identified rs12990970 (chr2:204,700,689) as a potential causative T1D SNP. In Treg cells, rs12990970 is found in a flanking active TSS (TssAFlnk) state and it is located within a Treg super-enhancer (Figure 2 & Additional file 1: Figure S2). This variant is located in a non-coding region between gene CTLA4 and CD28 and in past studies, and it has been associated with CTLA4 as it is an eQTL for CTLA4 expression in testis although not in T lymphocytes or whole blood (Additional file 3: Table S4) (11,13,45,46). Hi-C interaction signals however do not indicate that the rs12990970-containing region interacts with the CTLA4 promoter in Treg, instead Hi-C interactions indicates that this region form interactions with promoter and enhancer regions connected to the costimulatory receptor CD28 gene (Table 1 & Figure 2), a family member known to play a critical role in Treg homeostasis and function (51) suggesting CD28 is a novel target for this variant in Treg.

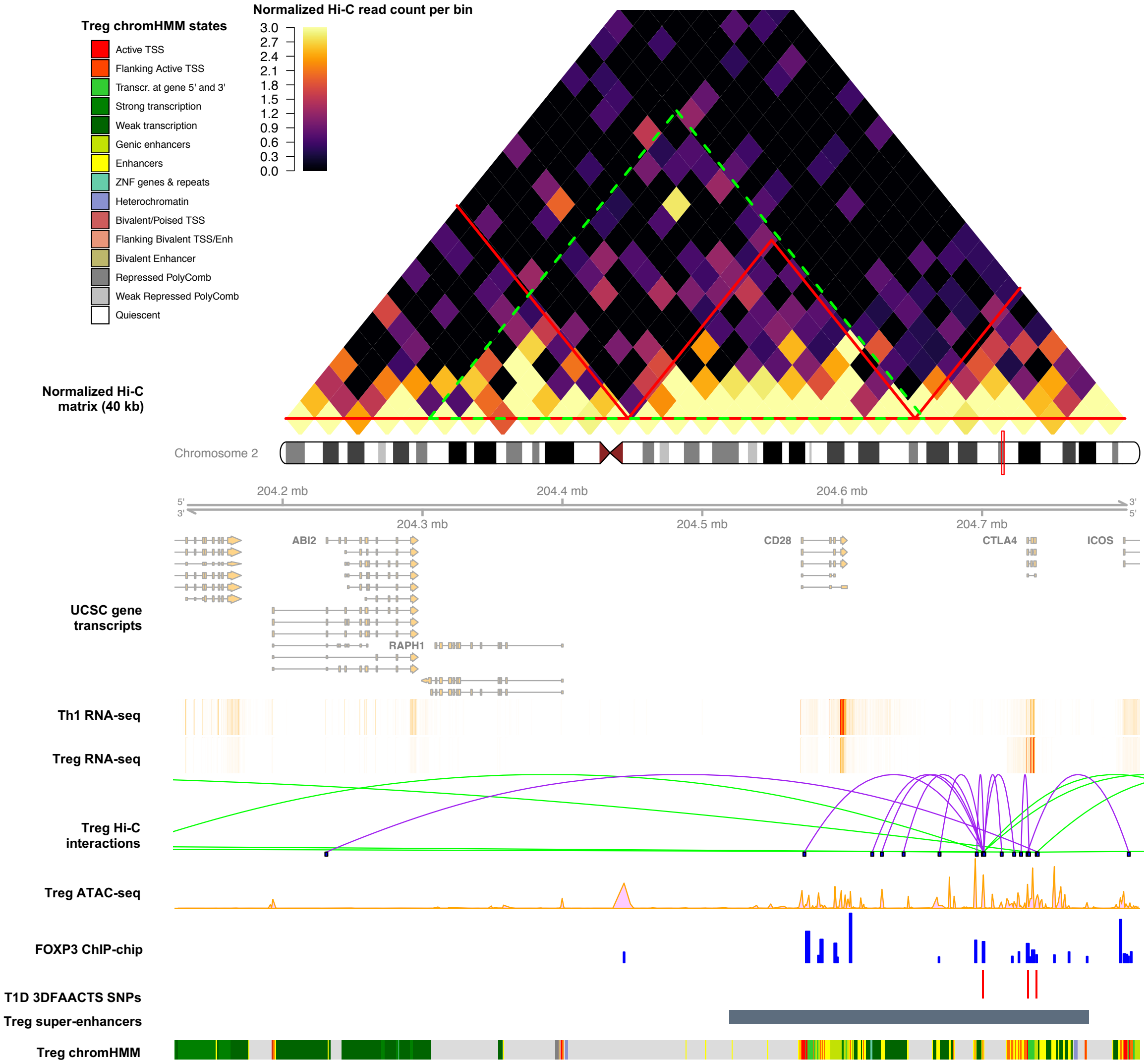


Figure 2: Visualisation of the CTLA4 region of filtered T1D SNPs on chromosome 2. Heatmap shows the Tregs Hi-C normalised interaction matrix (resolution of 40 kb) on chr2: 203922714-205092714. The red triangles indicate Topologically Associated Domains (TADs) and the large green-dotted triangle indicates the boundary of the current plot. Tracks displayed below the chromosome 2 ideogram display workflow datasets (filtered SNPs, FOXP3-binding sites and Treg ATAC-seq and Hi-C interactions) along with various types of cell type-specific data including UCSC Gene Transcript information, T cell subsets (Thelper1 and Treg) expression data, Treg super-enhancer sets and 15-state ChromHMM track. T1D 3DFAACTS SNPs within this region are rs12990970, rs231775 and rs3087243 (from left to right).

Another example is on chromosome 3, where Hi-C interactions indicated that the chemokine receptor genes CCR1, CCR2, CCR3 and CCR5 (Figure 3) are extensively linked in one TAD containing all of the filtered variants, indicating that these genes may be coordinately regulated. This is supported by previous RNA Pol-II ChIA-PET work (52) that detected interactions between chemokine gene clusters during immune responses including an increase in interactions amongst the CCR1, CCR2, CCR3, CCR5 and CCR9 genes during TNF stimulation of primary human endothelial cells (52) (Additional file 1: Figure S14). Recently, CCR2, CCR3 and CCR5 have been shown to have additional chemotaxis-independent effects on Treg cells with individual studies, reporting positive roles for individual chemokine receptors on CD25, STAT5, and FOXP3 expression and Treg potency (53–55), highlighting the importance of multiple genes at this locus on Treg function.

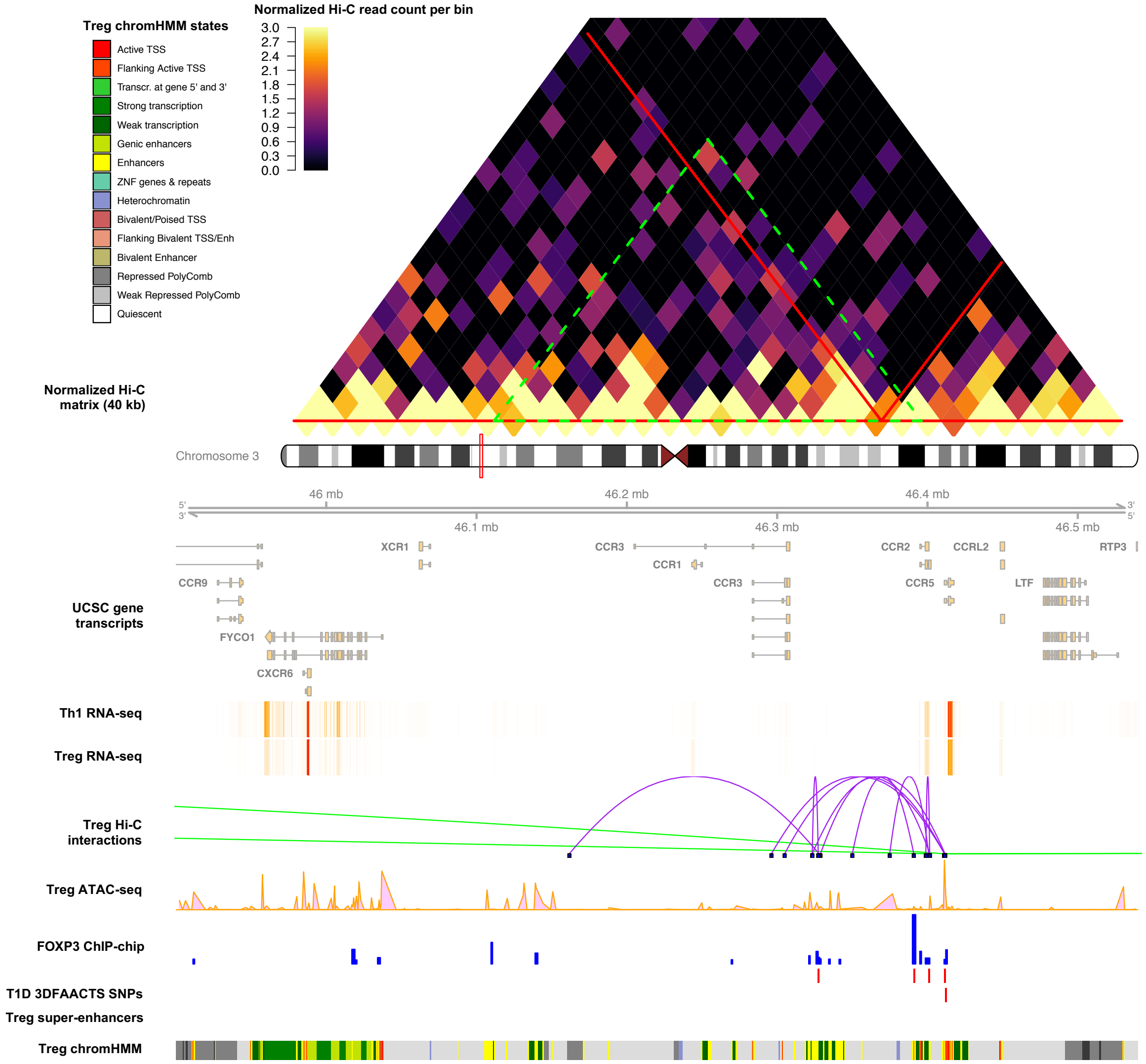


Figure 3: Visualisation of the CCR3/2/5 region of filtered T1D SNPs on chromosome 3. Heatmap shows the Tregs Hi-C normalised interaction matrix (resolution of 40 kb) on chr3: 45600000-46840000. The red triangles indicate Topologically Associated Domains (TADs) and the large green-dotted triangle indicates the boundary of the current plot. Tracks displayed below the chromosome 3 ideogram display workflow datasets (filtered SNPs, FOXP3-binding sites and Treg ATAC-seq and Hi-C interactions) along with various types of cell type-specific data including UCSC Gene Transcript information, T cell subsets (Thelper1 and Treg) expression data, Treg super-enhancer sets and 15-state ChromHMM track. T1D 3DFAACTS SNPs within this region are rs11718385, rs6441972, rs3138042, rs2856758 and rs1799988 (from left to right).

Filtered T1D variants are enriched at lineage specific T cell super-enhancers

SEs usually consist of a cluster of closely spaced enhancers that are defined by their exceptionally high level of transcription co-factor binding and enhancer-associated histone modifications (i.e. H3K27ac) compared to all other active enhancers within a specific cell type (56). SEs are also linked to the control of important processes such as cell lineage commitment, development and function (57). Analysing T cell SE information annotated in the Super-Enhancer Database (58) (SEdb; http://www.licpathway.net/sedb/), 8 out of the 14 variant-containing loci were found to contain filtered T1D variants located in SEs formed in various T cell lineages including Treg cells consistent with the enrichment of autoimmune-disease associated variants within T cell super enhancers reported previously (57) (Figure 4A). The loci containing the CTLA4 and CLEC16A genes were the only loci that overlapped with Treg-specific SEs. The existence of a Treg SE is consistent with the different regulation of CTLA4 in Treg cells compared with other T cell lineages (59) and a recent report linking T1D risk variants to altered CLEC16A expression in Treg (47). Five other SNPs are located within SEs in multiple T cell types including induced Treg (iTreg) suggesting the gene controlled by these SE play a broad role in T cell function. While no Treg SEs are detectable at the CD69 locus the T1D associated variants in this region overlapped with SEs formed in other T subsets. No T cell associated SEs are found in the loci containing the CCR1/2/3/5, PTPN2, RASGRP1 and HORMAD2 genes (Figure 4A).

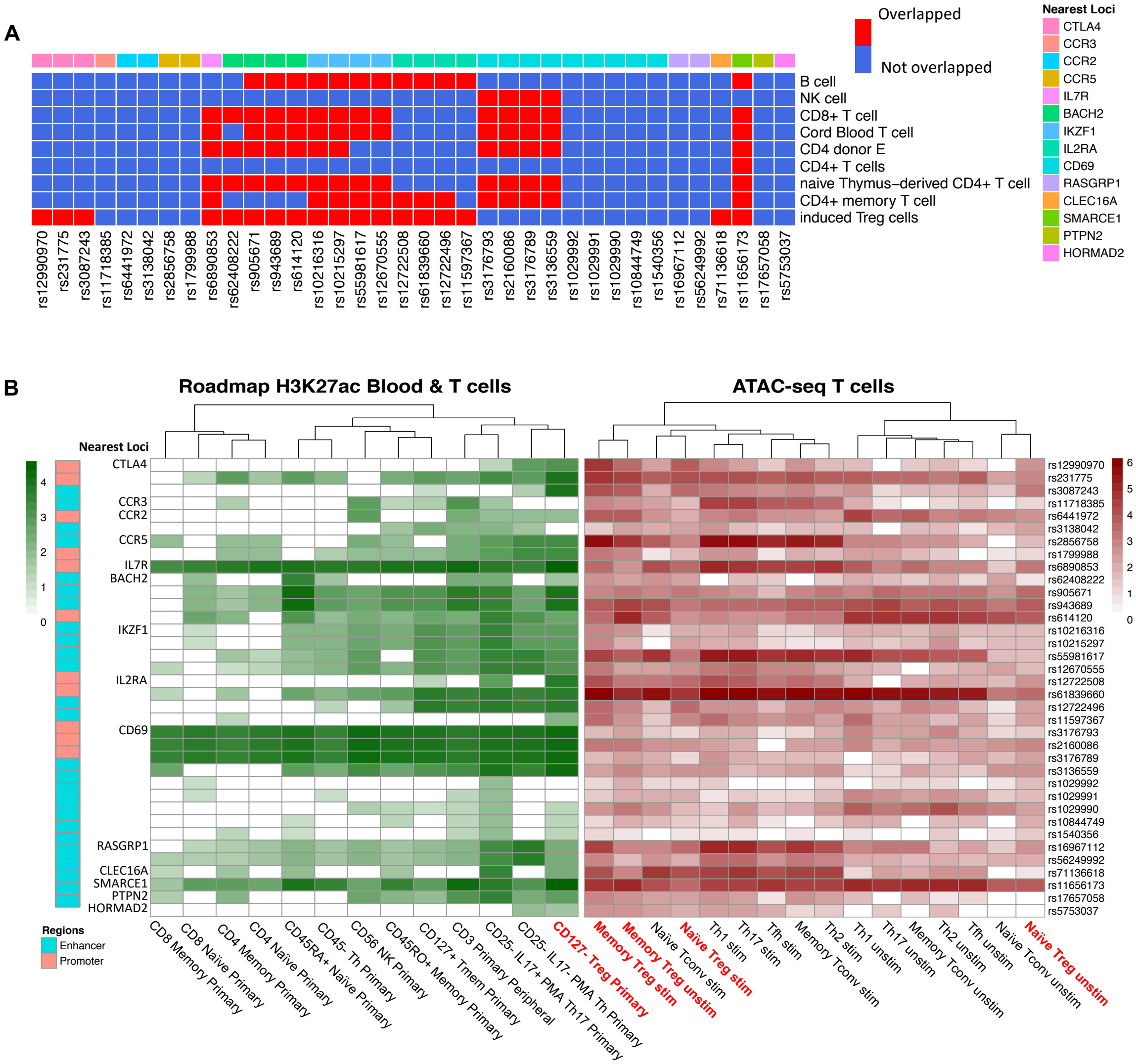


Figure 4: Integrating T1D 3DFAACTS SNPs with different data of T cell lineages. A. Heatmap showing overlapping status between T1D 3DFAACTS SNPs and super-enhancers of different T cell lineage from SEdb (58), where red indicates variants overlapping with SEs and blue indicates not overlapping. B. Enrichment of filtered T1D variants found within H3K27ac peaks from Epigenomics Roadmap and ATAC-seq peaks from multiple T cell lineages (34). Column names in red indicates Tregs specific datasets.

We then investigated the level of active enhancer marks (normalised H3K27ac-binding) and chromatin accessibility (normalised ATAC-seq peak coverage) overlapping each variant from Table 1 (Figure 4B). A range of tissue restriction patterns of chromatin states were observed using the NIH Epigenomics Roadmap data with enhancers displaying in general a more cell type-restricted pattern of H3K27ac signal compared to promoters. No variant was found to be located in a regulatory region that was exclusively active in Treg cells although rs12990970, rs231775 (CTLA4), rs11597367, rs12722508 (IL2RA) and rs5753037 (HORMAD2) are associated with a restricted H3K27ac pattern that included Treg. The absence of Treg-specific enhancers is consistent with FOXP3 binding data where FOXP3 binds many enhancer regions active in other T cell lineages to modify their activity in Treg cells (60). In particular, evidence suggests FOXP3 cooperates with other Thelper-lineage specifying transcription factors to diversify Treg cells into subsets that mirror the different Th-lineages (61–63). The majority of regions associated with the variants show an increase in chromatin accessibility upon stimulation in Treg and Thelper subsets consistent with increased enhancer activity upon T cell activation however in a few instances variants are located in regions that decrease in accessibility in stimulated Treg and Thelper subsets compared with their matched unstimulated counterpart. Notably these include the variants rs905671, rs943689 and rs614120 associated with BACH2. This is consistent with the reduction in BACH2 expression in CD4 T cells as they mature, and alteration to this repression is linked to proinflammatory effector function (64). Together these data are consistent with a model in which causal variants alter the output of enhancers that respond to environmental cues (65).

Filtered variants disrupt Transcription Factor Binding Sites (TFBS) including a FOXP3-like binding site

Fundamental to understanding the function of specific disease associated variants is the identification of the potential impact of these non-coding variants on transcription factor binding. Analysis of ATAC-seq datasets with HINT-ATAC (66), identified over 5 million active TF footprints in chromatin accessibility profiles from stimulated and resting Treg populations (Additional file 5: Table S6). By imposing the additional FOXP3 binding annotation to the footprint dataset, we identified 7 T1D-associated variants that have the potential to alter the binding of 9 TFs, suggesting the molecular mechanisms by which these variants could impact Treg function (Additional file 6: Table S7). Of these 7 SNPs, one SNP rs3176789 is located in an active TSS chromHMM state region, while the others are located either in enhancers or flanking active TSS that are associated with active enhancers, suggesting these variants might interrupt the binding of TFs to affect enhancer functions, with the potential for a network effect on multiple genes.

We then used GWAS4D (67), which computes log-odds of probabilities of the reference and alternative alleles of a variant for each selected TF motif to calculate binding affinity, to predict the regulatory effect of each variant (Supplemental Table 8). Several of the variants are predicted to alter the binding of transcription factors with known roles in Treg and other T cell lineages including Nuclear activator of T cells (NFATC2 & NFATC3, rs1029991) (68), interferon regulatory transcription factor (IRF, rs3176789) (69), myocyte enhancer factor 2 (MEF2, rs6441972 and rs3176789) and FOX (Forkhead box, rs614120) family members. In addition, variant (rs1029991) has the potential to alter the binding of YY1 recently identified as an essential looping factor involved in promoter-enhancer interactions (70). Other variants (rs1136618 and rs3176789) potentially alter the binding of the zinc finger protein ZNF384. Although expressed in T cells, the importance of ZNF384 in T cell biology has not yet been explored.

Of note, rs614120 is predicted to decrease the binding affinity of FOXA2 in this enhancer region (Additional file 6: Table S7). As FOXA2 is not expressed in the immune compartment, this SNP may interfere with the binding of another member of the forkhead class of DNA-binding proteins eg FOXP3, which is localised to this region based on our FOXP3 ChIP (Figure 5). This suggests that a model in which rs614120 impacts the expression level of BACH2 and/or AFG1L by altered binding of a FOX protein to this enhancer.

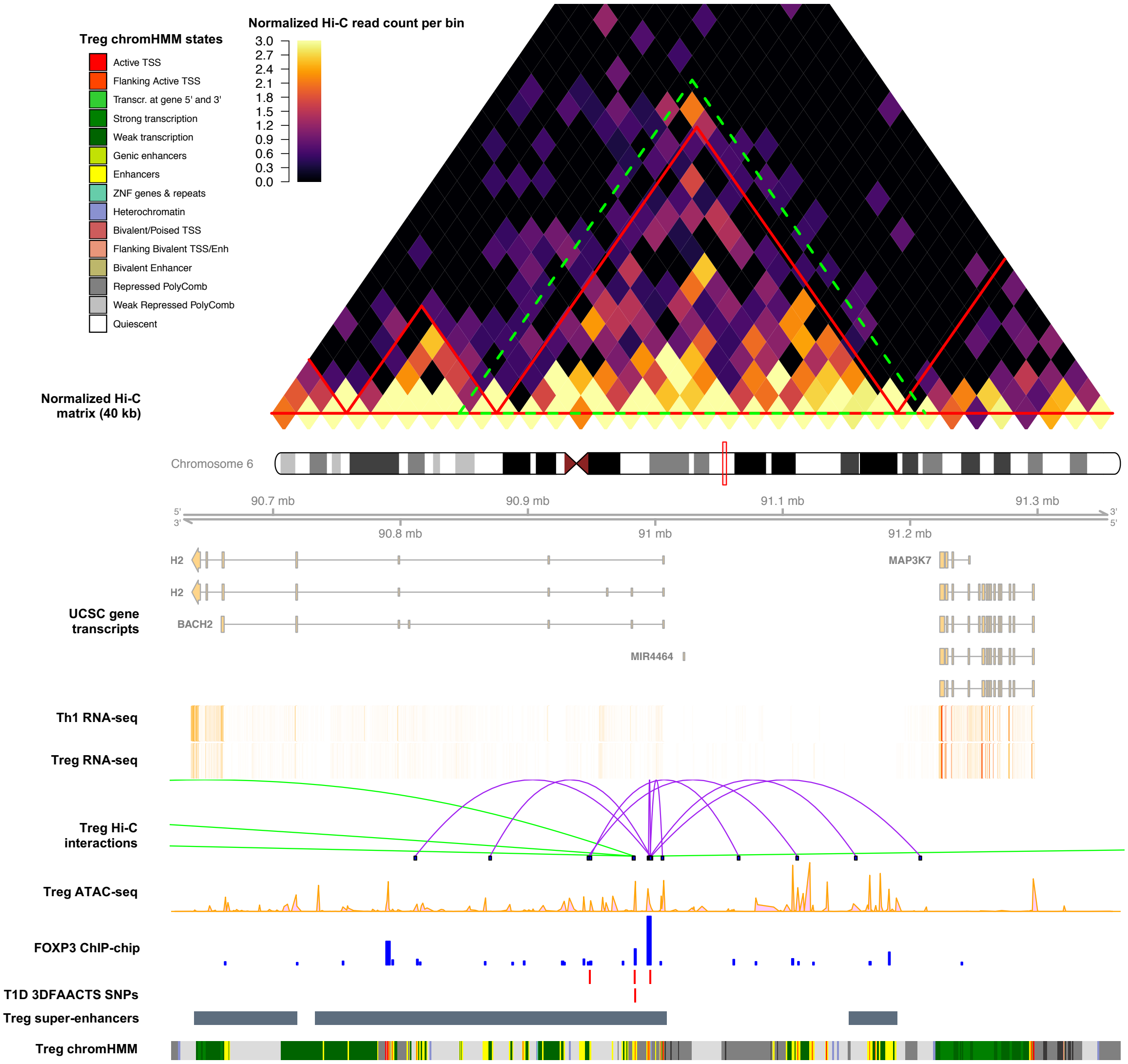


Figure 5: A. Visualisation of the BACH2 region of filtered T1D SNPs on chromosome 6. Heatmap shows the Tregs Hi-C normalised interaction matrix (resolution of 40 kb) on chr6: 90320000-91665000. The red triangles indicate Topologically Associated Domains (TADs) and the large green-dotted triangle indicates the boundary of the current plot. Tracks displayed below the chromosome 6 ideogram display workflow datasets (filtered SNPs, FOXP3-binding sites and Treg ATAC-seq and Hi-C interactions) along with various types of cell type-specific data including UCSC Gene Transcript information, T cell subsets (Thelper1 and Treg) expression data, Treg super-enhancer sets and 15-state ChromHMM track. T1D 3DFAACTS SNPs within this region are rs62408222, rs905671, rs943689 and rs614120 (from left to right).

Filtered variant rs1029991, is predicted to alter the binding of NFAT family members and/or YY1 to the enhancer region which has been linked to the cell surface expressed gene CD69. In addition, our analysis indicates that this enhancer also contacts the chromodomain helicase DNA-binding domain family member 4 (CHD4), indicating that CHD4 expression may be affected by this variant. Although CHD4 (Mi-2β) has not been previously linked to autoimmune diseases by GWAS studies, it has been shown to interacts with the T1D-associated genes IKZF1 and GATA3 and to play an important role in T cell development in the thymus and in T cell polarisation in the periphery including regulatory T cell subsets (71–73), consistent with altered regulation of CHD4 having the potential to contribute to T1D. Filtered variant rs3176789 is predicted to alter IRF and/or MEF2 binding linking these transcription factors to the regulation of the CD69, and CLEC family members CLECL1 and CLEC2D. The CD69 and CLEC2D genes have previously been associated with T1D by GWAS while CLECL1 has not. However, CLECL1 is a known target gene for eQTL rs3176789 (Additional file 3: Table S4), connecting this SNP and its associated regulatory region to CLECL1 expression rather than CD69 and suggesting a possible role for disrupted CLECL1 expression in Treg in T1D. Filtered variant rs6441972 is also predicted to influence the binding of MEF2 to a regulatory region in proximity to the promoter of CCR2. This region did not appear to interact with any other distal regulatory region or gene. Consistent with this variant disrupting CCR2 expression, CCR2 is a target gene for eQTL rs6441972, indicating that rs6441972 may result in altered CCR2 expression in a Treg in T1D by interfering with MEF2 binding.

Filtered Treg variants identified in other autoimmune diseases

The primary rationale of our filtering workflow is that autoimmune diseases like T1D are mediated by altered Treg functions. Hence, using GWAS data for other autoimmune diseases, we aimed to discover variants which potentially act by disrupting 3D gene regulation in Tregs. Similar to filtering fine-mapped T1D-associated SNPs, here we used the 3DFAACTS-SNP filtering workflow to process variants identified by Immunochip fine-mapping experiments and meta-analysis from three studies for a broad range of autoimmune and inflammatory diseases. SNPs associated with 10 autoimmune diseases were identified, representing 221 fine-mapped SNPs associated with multiple sclerosis (MS) (74); 69 SNPs identified by the meta-analysis of celiac disease (CeD), rheumatoid arthritis (RA), systemic sclerosis (SSc), and T1D(75) (which we refer to the 4AI dataset); and 244 SNPs identified by the meta-analysis of GWAS datasets for ankylosing spondylitis (AS), Crohn's disease (CD), psoriasis (PS), primary sclerosing cholangitis (PSC) and ulcerative colitis (UC)(76) (which we refer as 5ID dataset). Applying the 3DFAACTS-SNP pipeline we identified 9, 3 and 6 filtered variants from the MS, 4AI and 5ID datasets respectively (Additional file 7: Table S8). We identified putative target genes for these disease associated variants by Hi-C interactions resulting in 24, 8 and 8 genes linked to MS, 4AI and 5ID respectively (Additional file 7: Table S8). Many of these genes have either known roles in Treg differentiation, stability and function (GATA3 and CD84, ITCH, ILIRL2 and ILST) (77–85), or altered expression in human Treg in autoimmune-disease (ICA1, SESN3 and DLEU1) (86–88) and animal models of autoimmunity (SEPTIN7 and WWOX) (89).

Of the variants identified by 3DFAACTS-SNP, one variant (rs60600003) located at a locus on chromosome 7 was found to be associated with several diseases, including MS(74), celiac and systemic sclerosis(75), suggesting at least some of its interacting genes (ICA1, HERPUD2, SEPTIN7, ELMO1, DOCK4) may contribute to a common Treg defect in these diseases (Additional file 1: Figure S15 & Additional file 7: Table S8). When compared with the 36 variants identified from our T1D dataset analysis two variants, rs61839660 on chromosome 10 and rs3087243 on chromosome 2 were also prioritised by 3DFAACTS-SNP analysis of the 5ID and 4AI datasets respectively implicating their interacting genes SFMBT2 (rs61839660), ABI2 and IQCA1 (rs3087243) in the development of these diseases. While different variants were identified in the analysis of the various disease datasets, the regulatory elements in which these variants reside can be linked by Hi-C data to common candidate target gene such as PFKFB3 (rs12722496 and rs12722508 - T1D and rs947474 - 4AI). This is consistent with the view that common mechanistic pathways underlie some autoimmune diseases, although the specific risk allele within a locus can be disease-specific (90).

Similar to the filtered T1D SNPs, the GWAS filtered variants were more likely to be located within enhancer regions rather than promoters (Table 1 & Additional file 7: Table S8), surprising given that our defined enhancers cover less of the genome than promoters (enhancers: cover 2.23% of the human genome, while promoters cover 5.27%). This is also consistent with previous studies which have demonstrated an enrichment of disease associated variants at enhancer and super enhancer regions (57,91–93). We further annotated the filtered variants from these three datasets with GTEx eQTLs and Tregs eQTLs, identifying 4 SNPs that form an eQTL with a candidate gene target identified by Hi-C interactions (Additional file 7: Table S8). This included rs7731625-IL6ST and rs60600003-ELMO1, two SNP-gene contacts and eQTL pairings identified by 3DFAACTS-SNP as potential causative Treg defects in MS (rs7731625-IL6ST) and MS, T1D, celiac and systemic sclerosis (rs60600003-ELMO1), respectively. Of particular interest is the rs7731625-IL6ST pairing as IL6ST is a common signalling receptor of the IL6 family of cytokines known to have differing effects on Treg numbers and differentiation potential (83–85). Furthermore, the IL6-LIF axis has been proposed to regulate the balance of Th17/Treg cells with changes in Il6/LIF levels proposed to play a role in MS (82) highlighting a potential molecular mechanism for how the SNP variant rs7731625 may impinge on Treg function in MS.

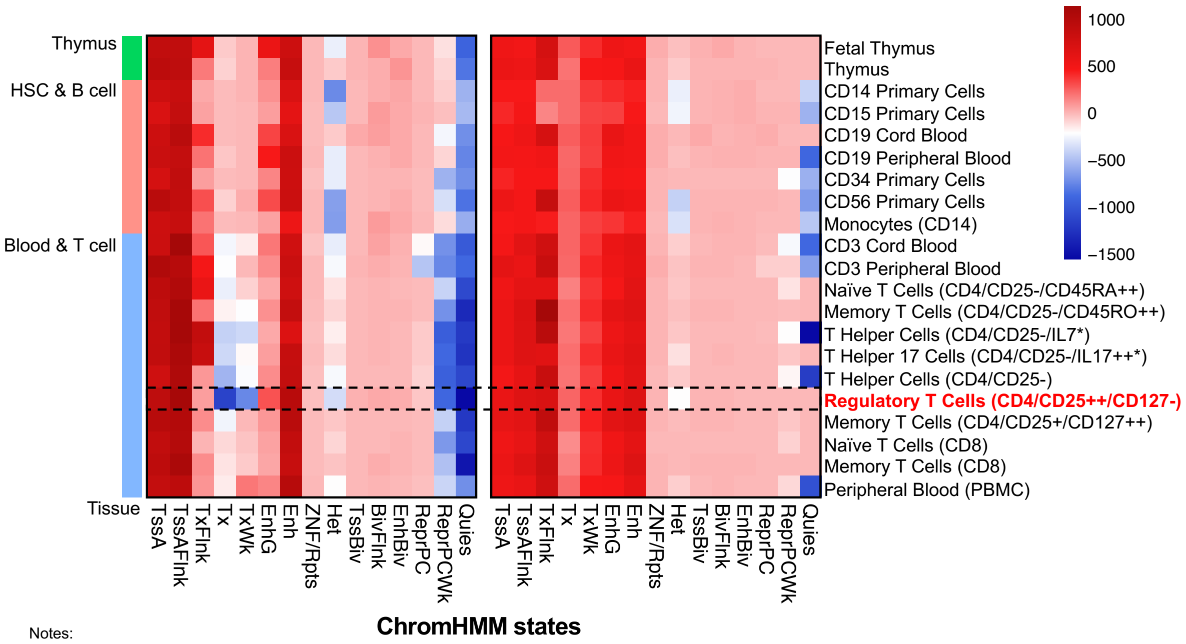
Identifying new variants that are candidates for impacting autoimmune disease

Most variants identified by GWAS have small effect sizes that together only represent a fraction of the heritability predicted by phenotype correlations between relatives (94). To account for this missing heritability, various models have been proposed including a highly polygenic architecture with small effect sizes of the causal variants (95,96), rare variants with large effect size (97,98) and epistatic mechanisms including gene-gene and gene-environment interactions (99,100). As a consequence many causal variants with small effect sizes are unlikely to reach genome wide significance in current GWAS whereas rare variants are often under-represented on SNP arrays (101). Lastly the preponderance of studies utilize populations of European descent which can result in a bias for SNPs with a higher minor allele frequencies in Europeans compared to other populations potentially limiting the relevance of these SNPs to the associated traits in non-Europeans (102). As an alternative approach to identify novel putative autoimmune disease-associated SNPs independently of association studies, we sampled 1,004,570 common variants (MAF > 0.1) from the Genome Aggregation Database (gnomAD) (version 3.0) (103) as inputs to our filtering workflow. Of these 808,857 overlapped with Tregs-specific Hi-C interactions, with 135,114 of these variants were located in promoter/enhancer regions and finally, 7,900 variants were located in FOXP3 binding regions (Additional file 8: Table S9). As a demonstration how this approach may complement current GWAS, 4,379 (55.7%) of the common variants we identified in gnomAD were not included in the largest GWAS T1D dataset to date (11) (Additional file 8: Table S9).

In order to further characterise the filtered gnomAD SNPs, we used *GIGGLE* (104) to compare the regions in which filtered SNPs reside against 15 predicted chromHMM genomic states across 127 cell types and tissues from Epigenomic Roadmap (34) (Figure 6 and Additional file 1: Figure S16), identifying positive and negative enrichment scores according to overlapping sets. Interestingly, although there was strong positive enrichment signal in active Tss (*TssA*), flanking active Tss (*TssAFlnk*) and enhancers (*Enh*) states in thymus, HSC, B- and T- cell groups, an enrichment was also observed across all cell types suggesting many of the enhancer and promoter regions and by extension their target genes are broadly expressed (Additional file 1: Figure S16). Moreover, unlike the 3DFAACTS-SNP analysis of GWAS derived data where filtered SNPs were enriched in enhancer regions, gnomAD derived SNPs are approximately evenly split between enhancers and promoter regions (Additional file 8: Table S9). Similarly, low/negative enrichment of the heterochromatin (Het) state was observed in all cell types whereas other inactive states such as repressed Polycomb (*ReprPC*, *ReprPCWk*) and the quiescent (*Quies*) states exhibited a negative enrichment in lymphoid cells. Interestingly, gnomAD SNPs demonstrated a strong negative enrichment in Treg cells for the chromatin states associated with strong transcription (*Tx*) and weak transcription (*TxWk*) potentially reflecting FOXP3 transcriptional repressor function (105).

Treg Hi-C data was used to explore the FOXP3-associated regulatory networks that include these SNPs in a Treg. For the regions identified to interact with the

7,900 variants located in FOXP3 binding regions by Hi-C we observed a strong positive enrichment of regulatory states such as *TssA*, *TssAFlnk*, *Tx*, *Txwk*, *EnhG* and *Enh* in blood, HSC, B and T cells, supporting a regulatory role for these interacting regions (Figure 6 and Additional file 1: Figure S17). In total 3,245 Treg expressed genes (mean FPKM > 1) (47) were found to be associated by Hi-C with variants identified by 3DFAACTS-SNP analysis of the common SNP gnomAD dataset. GO and Hallmark genes sets from the Molecular Signatures Database (MSigDB) (106,107) analysis of these 3,245 interacting Treg expressed genes were significantly enriched (adjusted P-value < 0.05) in relevant GO terms such as T cell activation and regulation of hematopoiesis (Additional file 1: Figure S18) and autoimmune/Tregs-related gene sets, including TNFα via NF-κB, IL6/JAK/STAT3, and IL2/STAT5 signaling pathways (Additional file 1: Figure S19). Integration of the filtered gnomAD variants with *cis* Treg eQTLs from the DICE database (46), further identified 943 common variants previously demonstrated to impact gene expression in Tregs (Additional file 8: Table S9). These 943 variants are connected by Hi-C interactions to 1038 genes in our analysis of which 121 (11.6%) form a *cis* eQTL pair with the 3DFAACTS identified SNPs. Importantly, interacting genes were significantly enriched (Fisher exact test, P value = 9.06e-24) in genes that are associated with 49 autoimmune diseases from GAAD (36) supporting the idea that we have identified potential novel disease associated molecular mechanisms.



Notes:

- indicates PMA and Ionomycin treatment.
- + and ++ indicates high and very high expression level
- - indicates no expression

Figure 6: Enrichment of 3DFAACTS gnomAD variants (left panel) and their interacting regions (right panel) found within NIH Epigenomics Roadmap samples. Enrichment test of filtered gnomAD SNPs against chromHMM states from 129 tissues and cell types from Epigenomics Roadmap using GIGGLE (104). Red coloured regions indicate positive enrichment of variants within cell-types and chromHMM states, while blue coloured regions indicate negative enrichment. Here we subset to enrichment in three tissue groups, including thymus, HSC & B cell and Blood & T cell, enrichment result of all samples can be found in Additional file 1: Figure S18 & 19.

We then integrated SNPs identified by 3DFAACTS-SNPs with the active TFBS dataset identified from Tregs ATAC-seq data by *HINT-ATAC* (66) (Additional file 8: Table S9) to identify potential molecular mechanisms of action of these non-coding SNPs. We found 870 filtered SNPs are located within active binding sites of 521 TFs indicating that they may impact TF binding. Accounting for the requirement of Treg expression of the TF (47) or its differential expression in Tregs compared to effector T cells (108), the number of variants with the potential to alter TF binding in a Treg was reduced to 693 and 108 variants respectively (Additional file 8: Table S9). Of the variants that potentially impact the binding of a TF expressed in a Treg, 19 were found to be an eQTL with its interacting gene partner identified by Hi-C. Included in this list were genes previously associated with Treg stability and viability, specific Treg subsets and pathways known to influence Treg differentiation and function. This is consistent with 3DFAACTS-SNP identifying potential novel variants that contribute to a Treg defect in disease. For example, Treg IL23R and FAS expression is associated with Treg/Th17 imbalances in IBD and the chronic inflammatory disease, acute coronary syndrome (109,110) and here using 3DFAACTS-SNP we predict rs1324551 and rs72676067 may contribute to this altered expression by disrupting the binding of the transcription factors RBPJ and POU2F2 respectively. Other genes are up-regulated in specific Treg subsets, including TBCID4 (follicular regulatory T cells) (111), ACTA2 (Placental bed uterine Tregs and tumour-infiltrating Treg) (48,112), and POLR1A (cold-exposed Brown adipose tissue Treg) (113), suggesting the identified common variants could lead to functional defects in these specific Treg subsets. A third group of genes have been shown to regulate growth factor signaling pathways that are known to influence Treg differentiation and function (114–116). In particular, we have identified variants that alter expression of the genes involved in TGF-β signaling (SPTBN1, CDC7 and SLC35F2) and WNT signaling (SPTBN1 and MCC). For example, 19 SNPs are linked to the SPTBN1 gene by Hi-C in our analysis, eight of which are identified as eQTLs with SPTBN1 in Tregs, and of these three (rs10170646, rs4455200 and rs13386146) overlap and potentially disrupt the binding of the transcription factors BCL6, HES2 and BATF-Jun heterodimer respectively prioritising these potential causative variants linked to allele-specific expression (ASE) of SPTBN1 in Treg. However further investigation is required to establish if altered SPTBN1 caused by these variants may contribute to any disease in response to TGF-β and WNT signaling pathways. Together, these data indicate that the 3DFAACTS SNP pipeline in combination with the gnomAD database has the potential to annotate novel disease associated variants and their potential molecular mechanisms of action, many of which have not previously been investigated in GWAS studies.

Discussion

GWAS and fine-mapping studies have identified over 50 candidate regions for T1D progression (11,13,117), however a broad understanding of the underlying disease mechanism has been difficult to elucidate without relevant functional information derived from cell-specific material. With the availability of whole genome annotation, we see that the majority of genetic risk lies in non-coding regions of the genome and is enriched in regulatory regions including promoters and enhancers. Traditionally, to understand how these variants may function they have been assigned to the nearest gene or genes within a defined linear distance. However this approach ignores the role of three-dimensional connectivity by which enhancers and repressors function to regulate transcription (118–120).

Recent approaches use statistical co-localization tests to link potential causal SNPs and quantitative trait loci (QTLs) to identify the genes regulated by GWAS loci (121). These methods require many samples in the correct cell type or physiological context and to date work best for local/*cis* QTLs, generally less than 1Mb in linear distance (118). An alternative approach used in this study and others (122,123) is to make use of chromosome conformation capture data to directly connect disease-associated regulatory regions to their target genes. As growing cellular and genomics evidence indicate that dysregulation of the Treg compartment contributes to autoimmune disease (27,124,125), we generated a cell type-specific 3D interaction profile in human regulatory T cells to establish an *in silico,* candidate loci reduction method to identify T1D-candidate regions that function in a Treg and the genes they affect. Open chromatin regions identified by ATAC-seq and regulatory regions identified by epigenetic marks such as histone H3K27ac can number in the tens of thousands in a specific cell type (47,126), we therefore initially focused on regulatory regions bound by the Treg-specific transcription factor FOXP3 given the essential role of FOXP3 in the Treg functional phenotype we hypothesized that candidate variants that are found within open, FOXP3-bound regions are likely to alter immunological tolerance. In addition as different autoimmune-diseases share genetic risk regions (41) we speculated that by identifying specific genetic variants that may contribute to T1D through the dysregulation of regulatory T cell functional fitness, this could be via mechanisms consistent across many autoimmune diseases (1,127,128).

The design and implementation of the 3DFAACTS-SNPs workflow champions a new data-centric view of functional genomics analysis, with the development of cell type-specific epigenomic and 3D datasets enabling researchers to narrow down on molecular changes at a fine-scale resolution. However, results shown in this study suggests that cell type-specific viewpoints can be broadened to a much more lineage (T cell) or immune (e.g. innate or adaptive) system-specific level. While we focused on Treg cells and expected to identify Treg-specific enhancer-controlled targets, based on the criteria of inclusion of FOXP3 binding data, no functional variant was uniquely accessible in only Tregs, nor were they specifically enriched with Treg-exclusive H3K27ac modified regions (Figure 4B). This likely reflects the propensity of FOXP3 to bind to enhancers active in multiple CD4+ T cell lineages (60) (Figure 4) to modify their output in a Treg-specific manner and therefore we cannot currently discern whether these filtered variants act predominantly in Tregs or on other CD4+ T cell subsets. The incorporation of context- and CD4+ T cell subset-specific gene expression (129) and epigenomic (123,130) data into the 3DFAACTS-SNPs workflow may help resolve this. Although we have focused here on using FOXP3-binding as a filtering criteria, it is known that other FOXP3-independent pathways are important for Treg function and the 3DFAACTS-SNPs workflow could be modified to incorporate other TFs or other epigenetic profiles such as CpG-demethylated regions (131) to further explore the relationship between disease-associated variants and these pathways.

In total using the 3DFAACTS-SNPs workflow we identified 36 novel candidate genes connected to variants in 12 T1D risk loci that could plausibly function in a Treg whereas we could not define plausible candidate Treg-specific activity at the other T1D risk regions that met all our filtering criteria. This may indicate that these other risk-regions are active in immune cell types other than a Treg or they impact genes and regulatory elements within a Treg that are not dependent upon FOXP3. As an example of how the 3DFAACTS-SNPs workflow can lead to testable insights into the molecular mechanisms of non-coding variants, the SNP rs614120 was found to be located in a FANTOM5 annotated T cell-specific enhancer region in the first intron of the BACH2 gene, and is predicted to disrupt the binding of Forkhead Transcription factor family member FOXA2 (Figure 5 and Additional file 6: Table S7). However, FOXA2 is not expressed in T cells, indicating that rs614120 might disrupt the binding of other Forkhead family members which bind to very similar DNA sequences, such as FOXP3, which is known to bind in this region (Figure 5). The 3DFAACTS-SNPs workflow further indicates that this enhancer region containing rs614120 interacts with the promoter of BACH2, forming a distal promoter-enhancer interactions, suggesting that rs614120 may disrupt FOXP3 binding to the enhancer leading to the dysregulation of BACH2 expression. It has been recently shown that Bach2 plays roles in the regulation of T cell receptor signalling in Tregs, including averting premature differentiation and assisting peripherally induced Treg development (132). Therefore, we suggested that this single variant may regulate BACH2 expression and ultimately may affect the progression of T1D, and this requires further experiments to verify. This can further aid the development of novel therapeutic approaches to restore function in Treg of patients with this genotype. This finding also suggests that variants can contribute to the causal mechanisms of disease by altering the efficacy/stability of TF binding in important regions such as enhancers or SEs. In the future, validation experiments such as ChIP or qPCR will be performed to validate the 3D relationships between genes and T1D variants discovered in this study.

The power of 3DFAACTS-SNPs is its ability to incorporate chromosome organisation in 3D and identify long-range interactions involving variant-containing regulatory regions leading to the identification of target genes that have not previously been associated with these diseases associated risk regions. This is illustrated by the finding that the majority (24/31) of Treg-expressed genes that interact with the T1D variants are not the closest gene in linear proximity and of these interacting genes 20 have not been previously associated with any autoimmune disease. For example, T1D 3DFAACTS SNP rs1029991 although located in linear proximity to the CD69 gene was found to contact the CHD4 gene (~3.2 mb away) (Additional file 3: Table S4) suggesting this variant is more likely to influence CHD4 expression than CD69. Interestingly, rs1029991 was not identified as a *cis*-eQTL for CHD4 in Tregs as it >3Mb away on the genome, with eQTLs being classified as cis when found <1Mb from their target gene. However, this interpretation using 3D information is largely dependent on the resolution of the available Hi-C data of the specific cell type. With low resolution Hi-C data (bin sizes larger than 10 kb), it may not be able to precisely conduct analyses to identify SNPs in regulatory regions.

The idea that high-order nuclear organisation coordinates transcription in times of immune challenge or tolerance was recently shown in a study demonstrating that 3D chromatin looping topology is important for a subset of long non-coding RNAs (lncRNAs), termed immune gene–priming lncRNAs (IPLs), to be correctly positioned at the promoters of innate genes (52). This positioning of the IPLs then allows for the recruitment of the WDR5–mixed lineage leukaemia protein 1 (MLL1) complex to these promoters to facilitate their H3K4me3 epigenetic priming (52). An example of long-range enhancer gene interactions in conveying autoimmune-disease risk in Treg cells has also recently been published (133). In this work a distal enhancer at the 11q13.5 locus associated with multiple autoimmune-disease risk, including T1D was found to participate in long-range interactions with the LRRC32 gene exclusively in Treg. Deletion of this enhancer in mice resulted in the specific loss of Lrcc32 expression in Treg cells and the inability of Treg to control gut-inflammation in an adoptive transfer colitis model. Furthermore CRISPR-activation experiments in human Tregs identified a regulatory element located in proximity to a risk variant rs11236797 that is capable of influencing LRRC32 expression. This data together highlights the mechanistic basis of how non-coding variants may function to interfere with Treg activity in disease. Although we did not identify this interaction in our final SNP-interaction list upon re-examination of our workflow this interaction was present in our Hi-C dataset, but it was filtered out as the enhancer is not bound by FOXP3. Coordinated genome topology has also been shown in immune cell lineage commitment, both at a loci (134,135) and compartment level (136), consistent with the concept of immune transcriptional “factories” where genes congregate in regions of the nucleus to undergo coordinated transcriptional activation (137).

Although a shared genetic aetiology between T1D and other immune-mediated diseases has been proposed we did not find a large overlap between the variants or interacting genes identified by 3DFAACTS SNP in T1D and other autoimmune disease datasets. The reason for this is not clear but may be a result of the relatively low number of input SNPs for the other autoimmune diseases. Irrespective of this, several candidate causal SNPs and genes including SFMBT2 (rs61839660), ABI2 and IQCA1 (rs3087243) and PFKFB3 (rs12722496 and rs12722508 - T1D and rs947474 - 4AI) were found to be common between T1D and other autoimmune diseases. Several of these genes such as SFMBT2, ABI2 and PFKFB3 have previously been implicated in the development of autoimmune diseases or play a role in critical T cell pathways suggesting these genes are likely targets that explain the molecular function of the risk variants. SFMBT2 is a methylated histone binding transcriptional repressor which has been associated with childhood onset asthma (138). ABI2 is required for actin polymerization at the T cell:APC contact site with loss of Abi1 in mice resulting in decreased TCR-mediated IL-2 production and proliferation (139). PFKFB3 is involved in both the synthesis and degradation of fructose-2,6-bisphosphate, a regulatory molecule that controls glycolysis in eukaryotes. Regulation of glycolysis has increasingly been implicated in shaping immune responses (140) and PFKFB3 has been associated with multiple autoimmune diseases (141). Importantly, reduced PFKFB3 enzyme activity leading to redox imbalance and apoptosis has been reported in CD4+ T from RA patients (142) directly linking the PFKFB3 gene to the disease.

A highly polygenic architecture with small effect sizes of many causal variants (95,96) has been proposed to account for missing heritability associated with phenotypic traits. Most of these small effect size variants have yet to be identified. Here we have begun to investigate whether common genetic variation found within populations could contribute to autoimmune diseases by altering gene-expression by altering enhancer and promoter output. In this study we illustrate this potential by accessing large population-scale variant resources in the gnomAD database, identifying 7,900 filtered common variants that have the potential to impact Treg function. Based on the search of discovered associations of autoimmune diseases (EFO_0005140) from the GWAS Catalog (143), over half of the variants surveyed here have not been used in large-scale autoimmune disease GWAS (11,76,144–150), precluding their assessment for potential disease risk in sampled disease/control populations. While filtered variants identified here are biased towards the inclusion of FOXP3-binding within the workflow, their potential immune response impact is highlighted by the finding that their interacting regions are positively enriched for transcription and enhancer -associated chromatin states (Figure 6, Additional file 1: Figure S16 & 17), eQTLs and potentially impacted TFBS (Additional file 8: Table S9). This potential accessibility of regulatory variants among a population could potentially explain additional variation in effector responses in T cell activation (151), relevant not only to autoimmune disease, but also to broader immune responses for example to SARS-CoV-2.

In conclusion, while we initially restricted the application of 3DFAACTS-SNP to Treg centric genome-wide interaction frequency profiles to give functional annotation in T1D data, we have demonstrated that valid interacting pairs from Hi-C dataset can be functionally mapped with high confidence from multiple disease datasets as well as whole genome variant datasets, which presents a valuable resource in establishing cell-type specific interactomes. Coupled with cell-type specific genomic data available from public repositories, such as the NIH Roadmap (34), Blueprint (152) and ENCODE (153) projects, this workflow provides a useful mechanism to identify potential mechanisms by which non-coding variants regulate disease causing genes, and identifies new targets for therapeutic modulation to treat or prevent disease.

Conclusion

Based on Treg ATAC-seq, Hi-C data, promoters and enhancers annotation and FOXP3 binding site, we developed a variant filtering workflow named 3DFAACTS-SNP to identify potential causative SNPs and their 3D interacting genes for T1D from GWAS fine-mapped variants. Our workflow can easily be used with variants associated with other autoimmune diseases or even large population-scale variants.

Methods

Cell preparation

Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood obtained from healthy human donors with informed consent at the Women's and Children's Hospital, Adelaide (ethics approval and consent see Declarations section). Cells were labelled with the following fluorochrome conjugated anti-human monoclonal antibodies: anti-CD4 (BD Biosciences, BUV395 Mouse Anti-Human), anti-CD25 (BD Biosciences, BV421), anti-CD127 (BD Biosciences, PE-CF594) and viability dye (BD Biosciences, BD Horizon Fixable Viability Stain 700) for FACS analysis by surface expression staining. Regulatory T (Treg) cells were sorted as CD4+ CD25hi CD127dim population (>90% purity). Following cell sorting Treg cells were plated at 100,000 cells per well in a 96-well U-bottom plate and maintained in complete X-VIVO 15 culture media (X-VIVO 15 Serum-free media supplemented with 2 mM HEPES pH 7.8, 2 mM L-glutamine and 5%

heat inactivated human serum) in 400U/mL rIL-2 for 2 hours at 37oC in a humidified 5% CO2 incubator prior to cell preparation for ATAC-seq experiment.

ATAC-seq library preparation and high-throughput sequencing

Treg cells were rested for 2-hour post sort and then were either left untreated or stimulated with beads conjugated with anti-CD3 and anti-CD28 antibodies (Dynabeads Human T-Expander CD3/CD28, Gibco no. 11141D, Life Technologies) in complete X-VIVO 15 culture in 400U/mL rIL-2 at a cell/bead ratio of 1:1 for 48 hours. After 48 hours Dynabeads were removed from culture medium by magnetic separation. Omni ATAC-seq was then performed as described previously (154) with minor modifications. Briefly, cells with 5-15% dead cells were pretreated with 200U/μL DNase (Worthington) for 30 minutes at 37^{o}C prior to ATAC-seq experiments. Treg cells (50,000) were lysed in 50μL of cold resuspension buffer (RSB: 10 mM Tris-HCl pH 7.4, 10 mM NaCl, and 3 mM MgCl2) containing 0.1% NP40, 0.1% Tween-20, and 0.01% digitonin on ice for 3 minutes. The reaction was then washed with 1mL of ATAC-seq RSB containing 0.1% Tween-20 by centrifugation at 500 xg for 10 minutes at 4^{o}C and the nuclei were resuspended in 50μL of transposition mix (30μL 2× TD buffer, 3.0μL Tn5 transposase, 16.5μL PBS, 0.5μL 1% digitonin and 0.5μL 10% Tween-20) (Illumina Inc). The transposition reaction was incubated at 37^{o}C for 45 minutes in a thermomixer with 1000 rpm mixing. The reaction was purified using a Zymo DNA Clean & Concentrator-5 (D4014) kit. All libraries were amplified for a total of 9 PCR cycles and size selection was carried out to enrich for a fragment size window of 200 to 900bp prior to sequencing. Libraries were quantified by PCR using a KAPA Library Quantification Kit for NGS (KAPA Biosystems, Roche Sequencing). Barcoded libraries were pooled and sequenced on a paired-end 75-cycle Illumina NextSeq 550 High-Output platform (Illumina) to an average read depth of 37.1 million reads (± 4 million) per sample.

Treg sample preparation, Hi-C library production and high-throughput sequencing

Cord blood was obtained with informed consent at the Women's and the Children's Hospital, Adelaide (HREC1596; WCHN Research Ethics Committee). Mononuclear cells were isolated from cord blood postpartum as previously described (155). Briefly, cord blood CD4^{+}CD25^{+}(Treg) were isolated from purified mononuclear cells using a Regulatory CD4^{+}CD25^{+}T Cell Kit (Dynabeads; Invitrogen, Carlsbad, CA). Ex vivo expansion of isolated T cell populations (1 × 10^{6} cells per well in a 24-well plate) were performed in X-Vivo 15 media supplemented with 5% human AB serum (Lonza, Walkersville, MD), 20 mM HEPES (pH 7.4), 2 mM L-glutamine, and 500 U/ml recombinant human IL-2 (R&D Systems, Minneapolis, MN) in the presence of CD3/CD28 T cell expander beads (Dynabeads; Invitrogen; catalogue no. 111-41D) at a bead-to-cell ratio of 3:1. Cell harvesting, Formaldehyde cross-linking (2%) and nuclei isolation was per (156,157). Treg cell nuclei were frozen in aliquots of 1x10^{7}. The in situ Hi-C procedure was carried out as per Rao et al, (2014) (158) with the following modifications MboI digestion was carried out in CutSmart® Buffer (NEB) and biotin-14-dCTP (Invitrogen; catalogue no. 19518018) replaced biotin-14-dATP in the reaction to end-fill MboI overhangs. This modification aims to reduce experiment cost and it was based on the methodology as per Naumova et al, (2012). To generate DNA suitable for library construction ligated DNA in TE buffer (10mM Tris-HCL, pH8.0 and 0.1mM EDTA, pH 8.0) was sheared to an average size of 300-500bp using a Covaris S220 (Covaris, Woburn, MA) instrument with the following parameters; 130ul in a microTube AFA fibre, 140 peak incidence power, 10% Duty cycle 10%, 200 cycles per burst for 55 seconds. Sheared fragment ends were made suitable for adapter ligation with a NEBNext® Ultra II End Repair/dA-Tailing Module (NEB #E7546). For adapter ligation the End Prep reaction was split into two and appropriately diluted NEBNext Adaptor ligated to fragment ends using the NEBNext Ultra II Ligation module. Hi-C libraries were split between 5 separate PCR reactions and directly amplified off the T1 beads using NEBNext Index Primers (set 1) and the NEBNext® Ultra™ II Q5® Master Mix. Library size distribution was determined using an Experion DNA 1K kit and library concentration estimated by real time qPCR using a Kapa universal Library quantitation kit (Roche Sequencing Solutions; 07960140001). Hi-C libraries were sequenced on a Illumina NextSeq 500 Mid-output platform (2x 150bp).

ATAC-seq data analysis

The sequencing data quality was determined using *FastQC* (ver. 0.11.7) (159) followed by trimming of Nextera adapters using *cutadapt* (ver. 1.14) (160).

Trimmed reads were aligned to the human hg19 (hs37d5) reference genome using *Bowtie2* (ver. 2.2.9) (161) with '-X 2000' setting. For each sample quality trimming was performed with option '-q 10' with unmapped and non-primary mapped reads filtered with option '-F 2828' using *Samtools* (ver. 1.3.1) (162). In this study, we used hg19 as reference genome instead of hg38 due to available annotation databases that mostly focus on hg19 information. Specifically in this work, the consistent genome build allows data to be comparable to hg19-mapped T1D SNPs data. PCR duplicates were then removed from Uniquely mapped paired reads using *Picard* (ver. 2.2.4). Mitochondrial reads, reads mapping to ENCODE hg19 blacklisted regions and mitochondrial blacklisted regions were filtered out using *BEDTools* (ver. 2.25.0). For peak calling the read start sites were adjusted to represent the center of Tn5 transposase binding event. Peaks were called from ATAC-seq data using *MACS2* (ver. 2.1.2) (163) and HINT-ATAC (66) was used to call footprints from the ATAC-seq peaks with parameters '--atac-seq --paired-end --organism=hg19'.

The peak summits from resting and stimulated Treg were concatenated and sorted by chromosome and then by position. The sorted peak summits were then handled using an in-house Python script *ATACseqCollapsing.py*, which adapted a peak processing approach described by Corces et al (154) to generate a list of non-redundant peaks. Briefly, through an iterative procedure, the peak summits are extended by 249 bp upstream and 250 bp downstream to a final width of 500 bp. Any adjacent peak that overlaps with the most significant peak (significance value defined by *MACS2*) within the interval is removed. This process iterates to the next peak interval resulting in a list of non-redundant significant peaks.

Hi-C data analysis

The raw sequencing read files were first processed using *AdapterRemoval* (ver 2.2.1a) (164) with default settings. The trimmed data were then analysed using *HiC-Pro* (ver 2.9.0) (32) with hg19 set as the reference genome and the GATCGATC as a potential ligation site. The valid interaction pairs of two technical replicates, which were stored in *allValidPairs* files were then concatenated into a single file followed by sorting based on the left interaction anchoring position of each interaction pair. The sorted interaction pairs were then processed using an in-house python script *allvalidpair2collapsingint.py* to generate non-redundant interactions. Similar to the merging process of ATAC-seq peaks, this is done by an iterative process, two anchor points of the first interaction pair are extended into windows with desired window sizes (in this case is 2 kb), the following interaction pair is removed only if both anchor points are within the previous interacting window, otherwise new interaction windows are generated, and the number of removed interaction pairs of each iteration are counted, resulting in non-redundant interaction pairs with window size of 2 kb. The merged interaction file was then processed using the functions *build_contact_map* and *ice_norm* from *HiC-Pro* to generate a normalised *n*n* matrix for subsequent visualisations.

RNA-seq data analysis

The raw sequencing data were first trimmed using *AdapterRemoval* (ver 2.2.1a) (164) with default parameters to remove sequencing adapters. Trimmed reads were then aligned to hg19 using *STAR* (ver 2.7.0d) (165). The resulting BAM files were converted into bedgraph files using *bamCoverage* from *deepTools* (166) with count normalised using counts per million mapped reads (CPM).

Topologically-associated domain identification

The valid interaction pairs of two technical replicates were concatenated together, followed by mapped to equal-size bins (40 kb) of the hg19 genome and normalised using *ICE* (32), resulting in a normalised interaction matrix. The matrix was then used as input to identify topologically-associated domains (TADs) via *TopDom* (167) with window size of 5.

Visualisation & Downstream analyses

Gene set enrichment analysis (GSEA) was performed using function *enrichr* from the R package *clusterProfiler* (168) with the hallmark gene sets from Molecular Signatures Database (MSigDB). Gene ontology (GO) analysis was performed using the R package *clusterProfiler* (168), with 0.01 as P-value threshold and 0.05 as adjusted P-value threshold (Benjamini-Hochberg adjusted). Visualisation of normalised Hi-C interaction matrices (Figure 2, 3, and 5, Additional file1: Figure S2-13) was performed on 40 kb resolution using an in-house R function *hicHeatmap*. The visualisations of individual filtered T1D-associated SNP loci

(Figure 2, 3, 5 and Additional file1: Figure S2-14) were constructed using the R packages *Gviz* (169), *GenomicInteractions* (170) and *coMET* (171). Visualisation of the GSEA network was performed using the R package *ggraph* (172).

Abbreviations

GWAS: Genome-wide association study

SNP: Single nucleotide polymorphism

T1D: Type 1 diabetes

Treg: Regulatory T cells

Tconv: Conventional T cells

FOXP3: Foxhead box protein 3

TF: Transcription factor

eQTL: Expression quantitative trait loci

ChIP-seq: Chromatin immunoprecipitation sequencing

ATAC-seq: Assay for transposase-accessible chromatin sequencing

Hi-C: High resolution chromosome conformation capture sequencing

TAD: Topologically-associated domain

SE: Super-enhancer

Declarations

Ethics approval and consent to participate

Cord blood used in this study was obtained with approval from the donor and the Children's, Youth and Women's Health Service Research Ethics Committee (HREC1596 and HREC 19/wchn/65 from the Women's and Children's Health Network Human Ethics committee). Buffy Coats were obtained from the Australian Red Cross (Material Supply Deed 19-03SA-02).

Consent for publication

Not applicable

Availability of data and materials

The Treg ATAC-seq and Hi-C datasets analysed during the current study are available in the European Nucleotide Archive (ENA) repository (PRJEB39882).

Published data:

FOXP3 ChIP-chip data used during the current study is available on Gene Expression Omnibus (accession no. GSE20995) (28).

Treg and Th1 RNA-seq data used during the current study is available on European Nucleotide Archive (accession no. ERR1198158 and ERR1198159) (48).

Database used in the current study:

NIH Roadmap Epigenomics Project: http://www.roadmapepigenomics.org/ (34)

SEdb: http://www.licpathway.net/sedb/ (58)

GAAD: http://gaad.medgenius.info/intro/ (36)

DICE: https://dice-database.org/landing (46)

gnomAD: https://gnomad.broadinstitute.org/ (103)

FANTOM5: https://fantom.gsc.riken.jp/5/ (38)

Source code for 3DFAACTS-SNP workflow and related in-house scripts are available in GitHub (https://github.com/ningbioinfostruggling/3DFAACTS-SNP).

Competing interests

Authors declare no competing interests.

Funding

This research was supported by a 2017 National Health and Medical Research Council (NHMRC) project grant (*#399123, #565314* and *1120543*). Dr Sadlon and Professor Barry are supported by a Women’s and Children’s Hospital Foundation Grant. Dr James Breen is supported by the James & Diana Ramsay Foundation.

Authors information

Ning Liu and Timothy Sadlon are joint first authors. Simon Barry and James Breen are joint corresponding authors

Affiliations

South Australian Health and Medical Research Institute, Adelaide, South Australia, 5000, Australia

Ning Liu & James Breen

Robinson Research Institute, University of Adelaide, Adelaide, South Australia, 5000, Australia

Ning Liu, James Breen, Timothy Sadlon, Ying Ying Wong & Simon Barry

Faculty of Health & Medical Sciences, University of Adelaide, Adelaide, South Australia, 5000, Australia

Ning Liu, Stephen M Pederson & James Breen

Womens & Childrens Hospital, North Adelaide, Adelaide, South Australia, 5006, Australia

Timothy Sadlon, Ying Ying Wong & Simon Barry

Contributions

All authors designed the study design. NL, TS, SB and JB wrote the manuscript, with additional help from YYW and SMP. NL developed and implemented all computational workflows and HiC-seq analyses. YYW analysed the ATAC-seq datasets and JB analysed public ChIP-seq and ATAC-seq data. TS and YYW constructed the ATAC-seq and Hi-C libraries.

Corresponding authors

Correspondence to James Breen and Simon Barry.

Acknowledgements

We are thankful for the access to publicly available resources, and grateful for discussions and comments from members of the Barry lab. We thank Dr R Gross for cell sorting expertise.

References

1. Long AS, Buckner JH. CD4+FOXP3+ T Regulatory Cells in Human Autoimmunity: More Than a Numbers Game. J Immunol. 2011;187:2061–6.

2. Atkinson MA, Eisenbarth GS, Michels AW. Type 1 diabetes. Lancet. 2014;383:69–82.

3. Fontenot JD, Gavin MA, Rudensky AY. Foxp3 programs the development and function of CD4+CD25+ regulatory T cells. Nat Immunol. 2003;4:ni904.

4. Bennett CL, Christie J, Ramsdell F, Brunkow ME, Ferguson PJ, Whitesell L, et al. The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3. Nat Genet. 2001;27:20–1.

5. Wildin RS, Ramsdell F, Peake J, Faravelli F, Casanova JL, Buist N, et al. X-linked neonatal diabetes mellitus, enteropathy and endocrinopathy syndrome is the human equivalent of mouse scurfy. Nat Genet. 2001;27:18–20.

6. Fontenot JD, Rudensky AY. A well adapted regulatory contrivance: regulatory T cell development and the forkhead family transcription factor Foxp3. Nat Immunol. 2005;6:ni1179.

7. Ono M. Control of regulatory T-cell differentiation and function by T-cell receptor signalling and Foxp3 transcription factor complexes. Immunology. 2020;160:24–37.

8. Sadlon T, Brown CY, Bandara V, Hope CM, Schjenken JE, Pederson SM, et al. Unravelling the molecular basis for regulatory T-cell plasticity and loss of function in disease. Clin Transl Immunology. 2018;7:e1011.

9. Hope CM, Welch J, Mohandas A, Pederson S, Hill D, Gundsambuu B, et al. Peptidase inhibitor 16 identifies a human regulatory T-cell subset with reduced FOXP3 expression over the first year of recent onset type 1 diabetes. Eur J Immunol. 2019;299:1057.

10. Redondo MJ, Yu L, Hawa M, Mackenzie T, Pyke DA, Eisenbarth GS, et al. Heterogeneity of type I diabetes: analysis of monozygotic twins in Great Britain and the United States. Diabetologia. 2001;44:354–62.

11. Bradfield JP, Qu H-QQ, Wang K, Zhang H, Sleiman PM, Kim CE, et al. A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. PLoS Genet. 2011;7:e1002293.

12. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther. 2011;13:101.

13. Onengut-Gumuscu S, Chen W-M, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat Genet. Nature Publishing Group; 2015;47:381–6.

14. Cerosaletti K, Schneider A, Schwedhelm K, Frank I, Tatum M, Wei S, et al. Multiple autoimmune-associated variants confer decreased IL-2R signaling in CD4+ CD25(hi) T cells of type 1 diabetic and multiple sclerosis patients. PLoS One. 2013;8:e83811.

15. Viglietta V, Baecher-Allan C, Weiner HL, Hafler DA. Loss of functional suppression by CD4+CD25+ regulatory T cells in patients with multiple sclerosis. J Exp Med. 2004;199:971–9.

16. Kim JM, Rasmussen JP, Rudensky AY. Regulatory T cells prevent catastrophic autoimmunity throughout the lifespan of mice. Nat Immunol. 2007;8:191–7.

17. Feuerer M, Shen Y, Littman DR, Benoist C, Mathis D. How punctual ablation of regulatory T cells unleashes an autoimmune lesion within the pancreatic islets. Immunity. 2009;31:654–64.

18. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput

Biol. 2012;8:e1002822.

19. Visscher PM, Wray NR, Zhang Q, Sklar P, I MM, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101:5–22.

20. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22:1748–59.

21. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nat Genet. 2007;39:1217–24.

22. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6:e1000888.

23. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45:1238–43.

24. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of cis regulatory variation in diverse human populations. PLoS Genet. 2012;8:e1002639.

25. Qian Y, Zhang L, Cai M, Li H, Xu H, Yang H, et al. The prostate cancer risk variant rs55958994 regulates multiple gene expression through extreme long-range chromatin interaction to control tumor progression. Sci Adv. 2019;5:eaaw6710.

26. 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;51:1442–9.

27. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet. 2013;45:124–30.

28. Sadlon TJ, Wilkinson BG, Pederson S, Brown CY, Bresatz S, Gargett T, et al. Genome-wide identification of human FOXP3 target genes in natural regulatory T cells. J Immunol. 2010;185:1071–81.

29. Krijger PHL, de Laat W. Regulation of disease-associated gene expression in the 3D genome. Nat Rev Mol Cell Biol. 2016;17:771–82.

30. Dekker J, Mirny L. The 3D genome as moderator of chromosomal communication. Cell. 2016;164:1110–21.

31. Gorkin DU, Leung D, Ren B. The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell. 2014;14:762–75.

32. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16:259.

33. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol.

2015;16:22.

34. Consortium, Roadmap Epigenomics, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature Publishing Group. Nature Publishing Group; 2015;518:317–30.

35. Beyer M, Thabet Y, Müller R-U, Sadlon T, Classen S, Lahl K, et al. Repression of the genome organizer SATB1 in regulatory T cells is required for suppressive function and inhibition of effector differentiation. Nat Immunol. 2011;12:898–907.

36. Lu G, Hao X, Chen W-H, Mu S. GAAD: A Gene and Autoimmiune Disease Association Database. Genomics Proteomics Bioinformatics. 2018;16:252–61.

37. Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet. 2013;14:661–73.

38. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.

39. Zacher B, Michel M, Schwalb B, Cramer P, Tresch A, Gagneur J. Accurate Promoter and Enhancer Identification in 127 ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN. PLoS One. 2017;12:e0169249.

40. Lowe CE, Cooper JD, Brusko T, Walker NM, Smyth DJ, Bailey R, et al. Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. Nat Genet. 2007;39:1074–82.

41. Suna O-G, Chen W-M, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat Genet. 2015;47:381–6.

42. Mlynarski WM, Placha GP, Wolkow PP, Bochenski JP, Warram JH, Krolewski AS. Risk of diabetic nephropathy in type 1 diabetes is associated with functional polymorphisms in RANTES receptor gene (CCR5): a sex-specific effect. Diabetes. 2005;54:3331–5.

43. Baker C, Chang L, Elsegood KA, Bishop AJ, Gannon DH, Narendran P, et al. Activated T cell subsets in human type 1 diabetes: evidence for expansion of the DR+ CD30+ subpopulation in new-onset disease. Clin Exp Immunol. 2007;147:472–82.

44. Marroquí L, Santin I, Dos Santos RS, Marselli L, Marchetti P, Eizirik DL. BACH2, a candidate risk gene for type 1 diabetes, regulates apoptosis in pancreatic β-cells via JNK1 modulation and crosstalk with the candidate gene PTPN2. Diabetes. 2014;63:2516–27.

45. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.

46. Schmiedel BJ, Singh D, Madrigal A, G V-GA, White BM, Jose Z-G, et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell. 2018;175:1701–15.e16.

47. Gao P, Uzun Y, He B, Salamati SE, Coffey JKM, Tsalikian E, et al. Risk variants disrupting enhancers of TH1 and TREG cells in type 1 diabetes. Proc Natl Acad Sci U S A. 2019;116:7581–90.

48. De Simone M, Arrigoni A, Rossetti G, Gruarin P, Ranzani V, Politano C, et al. Transcriptional Landscape of Human Tissue Lymphocytes Unveils Uniqueness of Tumor-Infiltrating T Regulatory Cells. Immunity. 2016;45:1135–47.

49. 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;485:376–80.

50. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods. 2015;72:65–75.

51. Bour-Jordan H, Bluestone JA. Regulating the regulators: costimulatory signals control the homeostasis and function of regulatory T cells. Immunol Rev. 2009;229:41–66.

52. Fanucchi S, Fok ET, Dalla E, Shibayama Y, Börner K, Chang EY, et al. Immune genes are primed for robust transcription by proximal long noncoding RNAs located in nuclear compartments. Nat Genet. 2019;51:138–50.

53. Zhan Y, Wang N, Vasanthakumar A, Zhang Y, Chopin M, Nutt SL, et al. CCR2 enhances CD25 expression by FoxP3+ regulatory T cells and regulates their abundance independently of chemotaxis and CCR2+ myeloid cells. Cell Mol Immunol. 2020;17:123–32.

54. Soler DC, Sugiyama H, Young AB, Massari JV, McCormick TS, Cooper KD. Psoriasis patients exhibit impairment of the high potency CCR5+ T regulatory cell subset. Clin Immunol. 2013;149:111–8.

55. Wang R, Huang K. CCL11 increases the proportion of CD4+CD25+Foxp3+ Treg cells and the production of IL-2 and TGF-β by CD4+ T cells via the STAT5 signaling pathway. Mol Med Rep. 2020;21:2522–32.

56. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153:307–19.

57. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013;155:934–47.

58. Jiang Y, Qian F, Bai X, Liu Y, Wang Q, Ai B, et al. SEdb: a comprehensive human super-enhancer database. Nucleic Acids Res. 2019;47:D235–43.

59. Walker L. Treg and CTLA-4: Two intertwining pathways to immune tolerance. J Autoimmun. 2013;45:49–57.

60. Samstein RM, Arvey A, Josefowicz SZ, Peng X, Reynolds A, Sandstrom R, et al. Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. Cell. 2012;151:153–66.

61. Chaudhry A, Rudra D, Treuting P, Samstein RM, Liang Y, Kas A, et al. CD4+ regulatory T cells control TH17 responses in a Stat3-dependent manner. Science. 2009;326:986–91.

62. Zheng Y, Chaudhry A, Kas A, deRoos P, Kim JM, Chu T-T, et al. Regulatory T-cell suppressor program co-opts transcription factor IRF4 to control T(H)2 responses. Nature. 2009;458:351–6.

63. Duhen T, Duhen R, Lanzavecchia A, Sallusto F, Campbell DJ. Functionally distinct subsets of human FOXP3+ Treg cells that phenotypically mirror effector Th cells. Blood. 2012;119:4430–40.

64. Tsukumo S-I, Unno M, Muto A, Takeuchi A, Kometani K, Kurosaki T, et al. Bach2 maintains T cells in a naive state by suppressing effector memory-related genes. Proc Natl Acad Sci U S A. 2013;110:10735–40.

65. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518:337–43.

66. Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription factor binding sites using ATAC-seq. Genome Biol. 2019;20:45.

67. Huang D, Yi X, Zhang S, Zheng Z, Wang P, Xuan C, et al. GWAS4D: multidimensional analysis of context-specific regulatory variant for human complex diseases and traits. Nucleic Acids Res. 2018;46:W114–20.

68. Wu Y, Borde M, Heissmeyer V, Feuerer M, Lapan AD, Stroud JC, et al. FOXP3 controls regulatory T cell function through cooperation with NFAT. Cell. 2006;126:375–87.

69. Fragale A, Gabriele L, Stellacci E, Borghi P, Perrotti E, Ilari R, et al. IFN regulatory factor-1 negatively regulates CD4+ CD25+ regulatory T cell differentiation by repressing Foxp3 expression. J Immunol. 2008;181:1673–82.

70. Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell. 2017;171:1573–88.e28.

71. Sridharan R, Smale ST. Predominant interaction of both Ikaros and Helios with the NuRD complex in immature thymocytes. J Biol Chem. 2007;282:30227–38.

72. Hosokawa H, Tanaka T, Suzuki Y, Iwamura C, Ohkubo S, Endoh K, et al. Functionally distinct Gata3/Chd4 complexes coordinately establish T helper 2 (Th2) cell identity [Internet]. Proceedings of the National Academy of Sciences. 2013. p. 4691–6. Available from: http://dx.doi.org/10.1073/pnas.1220865110

73. Shen E, Wang Q, Rabe H, Liu W, Cantor H, Leavenworth JW. Chromatin remodeling by the NuRD complex regulates development of follicular helper and regulatory T cells.

Proc Natl Acad Sci U S A. 2018;115:6780–5.

74. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science [Internet]. 2019;365. Available from: http://dx.doi.org/10.1126/science.aav7188

75. Márquez A, Kerick M, Zhernakova A, Gutierrez-Achury J, Chen W-M, Onengut-Gumuscu S, et al. Meta-analysis of Immunochip data of four autoimmune diseases reveals novel single-disease and cross-phenotype associations. Genome Med. 2018;10:97.

76. Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J, Han B, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet. 2016;48:510–8.

77. Wohlfert EA, Grainger JR, Bouladoux N, Konkel JE, Oldenhove G, Ribeiro CH, et al. GATA3 controls Foxp3+ regulatory T cell fate during inflammation in mice. J Clin Invest. The American Society for Clinical Investigation; 2011;121:4503–15.

78. Wang Y, Su MA, Wan YY. An essential role of the transcription factor GATA-3 for the function of regulatory T cells. Immunity. 2011;35:337–48.

79. Hua J, Davis SP, Hill JA, Yamagata T. Diverse Gene Expression in Human Regulatory T Cell Subsets Uncovers Connection between Regulatory T Cell Genes and Suppressive Function. J Immunol. 2015;195:3642–53.

80. Jin H-S, Park Y, Elly C, Liu Y-C. Itch expression by Treg cells controls Th2 inflammatory responses. J Clin Invest. 2013;123:4923–34.

81. Harusato A, Abo H, Ngo VL, Yi SW, Mitsutake K, Osuka S, et al. IL-36γ signaling controls the induced regulatory T cell–Th9 cell balance via NFκB activation and STAT transcription factors. Mucosal Immunol. 2017;10:1455–67.

82. Janssens K, Van den Haute C, Baekelandt V, Lucas S, van Horssen J, Somers V, et al. Leukemia inhibitory factor tips the immune balance towards regulatory T cells in multiple sclerosis. Brain Behav Immun. 2015;45:180–8.

83. Bin Dhuban K, Bartolucci S, d'Hennezel E, Piccirillo CA. Signaling Through gp130 Compromises Suppressive Function in Human FOXP3+ Regulatory T Cells. Front Immunol. 2019;10:1532.

84. Gao W, Thompson L, Zhou Q, Putheti P, Fahmy TM, Strom TB, et al. Treg versus Th17 lymphocyte lineages are cross-regulated by LIF versus IL-6. Cell Cycle. 2009;8:1444–50.

85. Hall AO, Beiting DP, Tato C, John B, Oldenhove G, Lombana CG, et al. The cytokines interleukin 27 and interferon-γ promote distinct Treg cell populations required to limit infection-induced pathology. Immunity. 2012;37:511–23.

86. Pesenacker AM, Chen V, Gillies J, Speake C, Marwaha AK, Sun A, et al. Treg gene signatures predict and measure type 1 diabetes trajectory. JCI Insight [Internet]. 2019;4. Available from: http://dx.doi.org/10.1172/jci.insight.123879

87. Jailwala P, Waukau J, Glisic S, Jana S, Ehlenbach S, Hessner M, et al. Apoptosis of CD4+ CD25(high) T cells in type 1 diabetes may be partially mediated by IL-2 deprivation. PLoS One. 2009;4:e6527.

88. Eiman MMA. Elucidating the molecular basis of multiple sclerosis and understanding the disease pathophysiology. Immunome Res. International Immunomics Society; 2016;12:1.

89. Schauer M, Kleinwort KJH, Degroote RL, Wiedemann C, Kremmer E, Hauck SM, et al. Interaction of septin 7 and DOCK8 in equine lymphocytes reveals novel insights into signaling pathways associated with autoimmunity. Sci Rep. 2018;8:12332.

90. Arakelyan A, Nersisyan L, Poghosyan D, Khondkaryan L, Hakobyan A, Löffler-Wirth H, et al. Autoimmunity and autoinflammation: A systems view on signaling pathway dysregulation profiles. PLoS One. 2017;12:e0187572.

91. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.

92. Corradin O, Scacheri PC. Enhancer variants: evaluating functions in common disease. Genome Med. 2014;6:85.

93. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.

94. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.

95. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42:565–9.

96. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169:1177–86.

97. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008;40:695–701.

98. Saint Pierre A, Génin E. How important are rare variants in common disease? Brief Funct Genomics. 2014;13:353–61.

99. Dempfle A, Scherag A, Hein R, Beckmann L, Chang-Claude J, Schäfer H. Gene–environment interactions for complex traits: definitions, methodological requirements and challenges. Eur J Hum Genet. 2008;16:1164–72.

100. Wei W-H, Hemani G, Haley CS. Detecting epistasis in human complex traits. Nat Rev Genet. 2014;15:722–33.

101. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. Genome Med. 2015;7:16.

102. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet. 2017;100:635–49.

103. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes [Internet]. bioRxiv. 2019 [cited 2019 Oct 21]. p. 531210. Available from: https://www.biorxiv.org/content/10.1101/531210v2

104. Layer RM, Pedersen BS, DiSera T, Marth GT, Gertz J, Quinlan AR. GIGGLE: a search engine for large-scale integrated genome analysis. Nat Methods. 2018;15:123–6.

105. Lopes JE, Torgerson TR, Schubert LA, Anover SD, Ocheltree EL, Ochs HD, et al. Analysis of FOXP3 reveals multiple domains required for its function as a transcriptional repressor. J Immunol. 2006;177:3133–42.

106. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.

107. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1:417–25.

108. Marek-Trzonkowska N, Piekarska K, Filipowicz N, Piotrowski A, Gucwa M, Vogt K, et al. Mild hypothermia provides Treg stability. Sci Rep. 2017;7:11915.

109. Ahern PP, Schiering C, Buonocore S, McGeachy MJ, Cua DJ, Maloy KJ, et al. Interleukin-23 drives intestinal inflammation through direct activity on T cells. Immunity. 2010;33:279–88.

110. Li Q, Wang Y, Wang Y, Zhou Q, Chen K, Wang YM, et al. Distinct different sensitivity of Treg and Th17 cells to Fas-mediated apoptosis signaling in patients with acute coronary syndrome. Int J Clin Exp Pathol. 2013;6:297–307.

111. Wing JB, Kitagawa Y, Locci M, Hume H, Tay C, Morita T, et al. A distinct subpopulation of CD25− T-follicular regulatory cells localizes in the germinal centers. Proc Natl Acad Sci U S A. National Academy of Sciences; 2017;114:E6400–9.

112. Wienke J, Brouwers L, van der Burg LM, Mokry M. Human regulatory T cells at the maternal-fetal interface show functional site-specific adaptation with tumor-infiltrating-like features. bioRxiv [Internet]. biorxiv.org; 2019; Available from: https://www.biorxiv.org/content/10.1101/820753v1.abstract

113. Medrikova D, Sijmonsma TP, Sowodniok K, Richards DM, Delacher M, Sticht C, et al. Brown adipose tissue harbors a distinct sub-population of regulatory T cells. PLoS One. 2015;10:e0118534.

114. Konkel JE, Zhang D, Zanvit P, Chia C, Zangarle-Murray T, Jin W, et al. Transforming Growth Factor-β Signaling in Regulatory T Cells Controls T Helper-17 Cells and Tissue-Specific Immune Responses. Immunity. 2017;46:660–74.

115. Chen W, Konkel JE. Development of thymic Foxp3+ regulatory T cells: TGF-β matters. Eur J Immunol. Wiley Online Library; 2015;45:958–65.

116. van Loosdregt J, Fleskens V, Tiemessen MM, Mokry M, van Boxtel R, Meerding J, et al. Canonical Wnt signaling negatively modulates regulatory T cell function. Immunity. 2013;39:298–310.

117. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet. 2009;41:703–7.

118. Tak YG, Farnham PJ. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin. 2015;8:57.

119. Hou L, Zhao H. A review of post-GWAS prioritization approaches. Front Genet. 2013;4:280.

120. Lee PH, Lee C, Li X, Wee B, Dwivedi T, Daly M. Principles and methods of in-silico prioritization of non-coding regulatory variants. Hum Genet. Springer Berlin Heidelberg; 2017;16:1–16.

121. Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front Genet. 2020;11:424.

122. Jeng MY, Mumbach MR, Granja JM, Satpathy AT, Chang HY, Chang ALS. Enhancer Connectome Nominates Target Genes of Inherited Risk Variants from Inflammatory Skin Disorders. J Invest Dermatol. 2019;139:605–14.

123. Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho S, et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat Genet. 2017;49:1602–12.

124. Fletcher JM, Lonergan R, Costelloe L, Kinsella K, Moran B, O'Farrelly C, et al. CD39+Foxp3+ regulatory T Cells suppress pathogenic Th17 cells and are impaired in multiple sclerosis. J Immunol. 2009;183:7602–10.

125. Lindley S, Dayan CM, Bishop A, Roep BO, Peakman M, Tree TIM. Defective suppressor function in CD4+ CD25+ T-cells from patients with type 1 diabetes. Diabetes. Am Diabetes Assoc; 2005;54:92–9.

126. Qu K, Zaba LC, Giresi PG, Li R, Longmire M, Kim YH, et al. Individuality and variation of personal regulomes in primary human T cells. Cell Syst. 2015;1:51–61.

127. Cvetanovich GL, Hafler DA. Human regulatory T cells in autoimmune diseases. Curr Opin Immunol. 2010;22:753–60.

128. Dominguez-Villar M, Hafler DA. Regulatory T cells in autoimmune disease. Nat Immunol. 2018;19:665–73.

129. Höllbacher B, Duhen T, Motley S, Klicznik MM, Gratz IK, Campbell DJ. Transcriptomic profiling of human effector and regulatory T cell subsets identifies predictive population signatures [Internet]. 2020 [cited 2020 Aug 21]. p.

2020.05.13.093567. Available from:
https://www.biorxiv.org/content/10.1101/2020.05.13.093567v1.abstract

130. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat Biotechnol. 2019;37:925–36.

131. Ohkura N, Yasumizu Y, Kitagawa Y, Tanaka A, Nakamura Y, Motooka D, et al. Regulatory T Cell-Specific Epigenomic Region Variants Are a Key Determinant of Susceptibility to Common Autoimmune Diseases. Immunity. 2020;52:1119–32.e4.

132. Sidwell T, Liao Y, Garnham AL, Vasanthakumar A, Gloury R, Blume J, et al. Attenuation of TCR-induced transcription by Bach2 controls regulatory T cell differentiation and homeostasis [Internet]. Nature Communications. 2020. Available from: http://dx.doi.org/10.1038/s41467-019-14112-2

133. Nasrallah R, Imianowski CJ, Bossini-Castillo L, Grant FM, Dogan M, Placek L, et al. A distal enhancer at risk locus 11q13.5 promotes suppression of colitis by Treg cells. Nature. 2020;583:447–52.

134. van Schoonhoven A, Huylebroeck D, Hendriks RW, Stadhouders R. 3D genome organization during lymphocyte development and activation. Brief Funct Genomics. 2020;19:71–82.

135. Duan J, Shi J, Fiorentino A, Leites C, Chen X, Moy W, et al. A rare functional noncoding variant at the GWAS-implicated MIR137/MIR2682 locus might confer risk to schizophrenia and bipolar disorder. Am J Hum Genet. 2014;95:744–53.

136. Isoda T, Moore AJ, He Z, Chandra V, Aida M, Denholtz M, et al. Non-coding Transcription Instructs Chromatin Folding and Compartmentalization to Dictate Enhancer-Promoter Communication and T Cell Fate. Cell. 2017;171:103–19.e18.

137. Papantonis A, Kohro T, Baboo S, Larkin JD, Deng B, Short P, et al. TNFα signals through specialized factories where responsive coding and miRNA genes are transcribed. EMBO J. 2012;31:4404–14.

138. Pividori M, Schoettler N, Nicolae DL, Ober C, Im HK. Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. Lancet Respir Med. 2019;7:509–22.

139. Zipfel PA, Bunnell SC, Witherow DS, Gu JJ, Chislock EM, Ring C, et al. Role for the Abi/wave protein complex in T cell receptor-mediated proliferation and cytoskeletal remodeling. Curr Biol. 2006;16:35–46.

140. Ganeshan K, Chawla A. Metabolic regulation of immune responses. Annu Rev Immunol. 2014;32:609–34.

141. Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong C, Fumis L, Karamanis N, et al. Open Targets Platform: new developments and updates two years on. Nucleic Acids Res. 2019;47:D1056–65.

142. Yang Z, Fujii H, Mohan SV, Goronzy JJ, Weyand CM. Phosphofructokinase

deficiency impairs ATP generation, autophagy, and redox balance in rheumatoid arthritis T cells. J Exp Med. 2013;210:2119–34.

143. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D1005–12.

144. Grant SFA, Qu H-Q, Bradfield JP, Marchand L, Kim CE, Glessner JT, et al. Follow-up analysis of genome-wide association data identifies novel loci for type 1 diabetes. Diabetes. 2009;58:290–5.

145. Zhu M, Xu K, Chen Y, Gu Y, Zhang M, Luo F, et al. Identification of Novel T1D Risk Loci and Their Association With Age and Islet Function at Diagnosis in Autoantibody-Positive T1D Individuals: Based on a Two-Stage Genome-Wide Association Study. Diabetes Care. 2019;42:1414–21.

146. International Multiple Sclerosis Genetics Consortium (IMSGC), Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet. 2013;45:1353–60.

147. Andlauer TFM, Buck D, Antony G, Bayas A, Bechmann L, Berthele A, et al. Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. Sci Adv. 2016;2:e1501678.

148. Huang C, Haritunians T, Okou DT, Cutler DJ, Zwick ME, Taylor KD, et al. Characterization of genetic loci that affect susceptibility to inflammatory bowel diseases in African Americans. Gastroenterology. 2015;149:1575–86.

149. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet. 2017;49:256–61.

150. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet. 2011;43:1193–201.

151. Cano-Gamez E, Soskic B, Roumeliotis TI, So E, Smyth DJ, Baldrighi M, et al. Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4+ T cells to cytokines. Nat Commun. 2020;11:1801.

152. Fernández JM, de la Torre V, Richardson D, Royo R, Puiggròs M, Moncunill V, et al. The BLUEPRINT Data Analysis Portal. Cell Syst. 2016;3:491–5.e5.

153. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46:D794–801.

154. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods. 2017;14:959–62.

155. Bresatz S, Sadlon T, Millard D, Zola H, Barry SC. Isolation, propagation and characterization of cord blood derived CD4+ CD25+ regulatory T cells. J Immunol Methods. 2007;327:53–62.

156. van de Werken HJG, de Vree PJP, Splinter E, Holwerda SJB, Klous P, de Wit E, et al. 4C technology: protocols and data analysis. Methods Enzymol. 2012;513:89–112.

157. Splinter E, de Wit E, van de Werken HJG, Klous P, de Laat W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. Methods. 2012;58:221–30.

158. Rao SSP, 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;159:1665–80.

159. Andrews S, Others. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

160. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. journal.embnet.org; 2011;17:10–2.

161. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. nature.com; 2012;9:357–9.

162. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. academic.oup.com; 2009;25:2078–9.

163. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. genomebiology.biomedcentral.com; 2008;9:R137.

164. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016;9:88.

165. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

166. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42:W187–91.

167. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res. 2016;44:e70.

168. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.

169. Hahne F, Ivanek R. Visualizing Genomic Data Using Gviz and Bioconductor. Methods Mol Biol. 2016;1418:335–51.

170. Harmston N, Ing-Simmons E, Perry M, Barešić A, Lenhard B. GenomicInteractions: An R/Bioconductor package for manipulating and investigating chromatin interaction

data. BMC Genomics. 2015;16:963.

171. Martin TC, Yet I, Tsai P-C, Bell JT. coMET: visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns. BMC Bioinformatics. 2015;16:131.

172. Pedersen TL, Pedersen MTL, LazyData T, Rcpp I, Rcpp L. Package “ggraph.” Retrieved January. cran.uni-muenster.de; 2017;1:2018.

173. Corradin O, Scacheri PC. Enhancer variants: evaluating functions in common disease. Genome Med. 2014;6:85.

Supplementary information

Note: Supplementary Tables and Figures are hosted on Figshare

(https://figshare.com/s/f4ab4f81e0963914d7fd).

Additional file 1.

Figures. S1–19, and Tables S1-2.

Additional file 2: Table S3.

Promoters and enhancers used in the 3DFAACTS-SNP workflow in this study.

Additional file 3: Table S4.

T1D 3DFAACTS SNPs identified using 3DFAACTS-SNP workflow from T1D fine-mapped SNPs and their 3D interacting genes.

Additional file 4: Table S5.

Topologically-associated domains (TADs) identified using TopDom from Treg Hi-C data and the relationship between Hi-C interactions of T1D 3DFAACTS SNPs and identified TADs.

Additional file 5: Table S6.

Transcription factor footprint identified from rest and stimulated Tregs ATAC-seq data using HINT-ATAC.

Additional file 6: Table S7.

T1D 3DFAACTS SNPs that are located in active transcription factor footprint in Tregs. And the binding affinity effect of these SNPs with their overlapped transcription factor calculated using GWAS4D.

Additional file 7: Table S8.

3DFAACTS SNPs identified from fine-mapped and meta-analysis SNP datasets from 3 published studies.

Additional file 8: Table S9.

3DFAACTS SNPs identified from gnomAD common (MAF >= 0.1) SNPs.

Chapter 4

Landscape of statistically significant chromatin interaction profiles of cell lines and primary tissues in the human genome

Landscape of statistically significant chromatin interaction profiles of cell lines and primary tissues in the human genome

Ning Liu[1,2,3], Hamid Alinejad-Rokny[4*], James Breen[1,2,3,5*]

[1] South Australian Health & Medical Research Institute, Adelaide, Australia
[2] Robinson Research Institute, University of Adelaide, Adelaide, Australia
[3] Adelaide Medical School, University of Adelaide, Adelaide, Australia
[4] Systems Biology and Health Data Analytics Lab, The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, NSW, 2052, Australia
[5] South Australian Genomics Centre (SAGC), Adelaide, Australia

* corresponding authors

Contributions: NL and JB developed and conceived the study in collaboration with HAR. NL wrote and developed approximately 90% of the manuscript, with editing provided by HAR and JB. Manuscript is formatted for submission to *Genome Biology*.

Abstract

Background

High resolution chromosome conformation capture (Hi-C) sequencing has been used in many studies to identify chromatin interactions and uncover 3D chromosome structures across the genome. Additionally, chromatin interactions have been shown to play an important functional role in gene regulation and maintaining spatial organisation, whereby physical connections between specific regions of the genome are maintained to enable transcriptional processes to be conducted. While cell-lines and tissues have been assessed for interactions at an individual sample level, a full meta-analysis of available HiC datasets have yet to be produced that determines statistically significant interactions that take place across all samples.

Results

Taking advantage of 348 publicly available Hi-C datasets, we developed a comprehensive analysis workflow that aims to identify significant interactions using a statistical algorithm MaxHiC, identifying MaxHiC-detected statistically significant interaction (MADSSI) profiles for available datasets. After filtering for technical biases down to 173 samples were analysed across 51 cell lines and tissues, we found that unique, tissue/cell-line-specific interactions (i.e. found in only one tissue) made up 62.3% of all MADSSI profiles, tended to contact type-specific genes and were likely to contact regions over longer distances than non-unqiue interactions. Non-unique interactions on the other hand were more likely to contact genes that are active and expressed in multiple cells and tissues. Finally, we characterised 2,442 interaction “hot zones”, regions observing statistically significant interactions in over a majority of analysed tissue/cell-types. We found hot zones were significantly enriched for CTCF-binding sites, an important structural transcription factor, and located close to topologically-associated domains (TADs) boundaries. Furthermore, hot zones are also enriched for active histone markers such as H3K27ac and H3K4me1, enhancer-like signatures of candidate *cis* regulatory elements (cCREs), demonstrating their structural importance in maintaining chromosome structure and driving gene regulation.

Conclusion

Altogether, our catalogued profiles provide another valuable resource for researchers to enable the identification of functional interactions for specific cell/tissue types or diseases.

Keywords

Chromatin interactions; Hi-C; 3D chromosome structure; Statistically significant interactions.

Background

In any eukaryotic genome such as the human genome, chromosomes are packaged into the nucleus in a hierarchical structure, which can bring linearly distal DNA segments into close proximity, resulting in long-range chromatin interactions [1]. Incorporated with high throughput sequencing technologies, the high resolution chromosome conformation capture (Hi-C) assay was designed to facilitate the investigation of the three dimensional (3D) architecture of a genome and has become one of the most popular approaches to identify 3D interactions and construct contact maps across the genome [2–4]. Hi-C uses formaldehyde and enzymes such as restriction enzymes and DNase I to capture cross-linked DNA fragments and to construct a Hi-C DNA library, where each fragment contains two pieces of cross-linked DNA that are joined together via ligating enzymes [2]. Sequencing the ligated DNA from the Hi-C library using paired-end technology, enables the detection of DNA regions that physically interact with each other, potentially across long distances.

With the reduction of the sequencing cost in the recent decade [5], more and more groups have published their own Hi-C data, providing a large amount of accessible data that can be used to categorise cell line/tissue-specific Hi-C interactions. However, the complexity and scale of information available, as well as the quality of data available, can cause issues for researchers wanting to use data to identify physical interactions in other domains. Key to this issue is the identification of Hi-C interactions that are significantly enriched in a HiC profile relative to statistical noise, owing to their higher biological importance [6,7]. Current statistical models can be used to account for random polymer looping and other types of noise, which can bias downstream analyses using Hi-C interactions [6]. It has also been shown that the statistical estimation of Hi-C interactions can assist in identifying potentially functional links between regions like promoters and enhancers [6–9].

One of the most important applications of chromatin interactions is using its novel linkage information to bridge non-gene-coding regions or elements to genes. Single nucleotide polymorphisms (SNPs) have been detected by genome-wide association studies (GWAS) to discover variations that are associated with diseases or important traits. However, the majority of diseases-associated SNPs are found in non-gene-coding regions [10], making them difficult to be interpreted. Integrative analyses with genetic variation and gene expression, namely expression quantitative trait loci (eQTLs), have helped determine target genes for variants located within non-gene-coding regions [11,12], however the underlying mechanism of how they impact the expression level of their target genes is still unknown. Using chromatin interactions, it is able to bring different non-gene-coding genomic elements such as gene promoters, enhancers, transcription factor binding sites, SNPs and eQTLs together to construct the complex gene regulation networks governed by 3D genome structure and functionally interpret non-gene-coding elements [13–15].

In a previous study, we developed MaxHiC, a method that uses the adaptive moment estimation (Adam) algorithm [16] to maximize the likelihood of the observed Hi-C interactions and is able to identify statistically significant interactions based on a negative binomial distribution background model. More importantly, MaxHiC significantly outperformed existing Hi-C models in identifying biologically relevant interactions [9]. In this study, we aimed to collect all published Hi-C data of human cell lines and tissues and generate a comprehensive map of statistically significant interaction profiles in 51 cell lines and tissues. These profiles can be used to interpret genomic variants such as traits/diseases-associated GWAS SNPs and quantitative trait loci (QTLs), reveal regulations governed by 3D genome structure such as promoter-enhancer interactions, identify novel connections from genes important for diseases.

Results

Data selection and process

Our study was carried out in three stages: data collection, data processing and profile generation (Figure 1). Firstly, we collated all published Hi-C datasets from untreated or control samples of human cell lines and tissues in the European Nucleotide Archive [17] and 4DN data portal [18]. We specifically gathered data generated by Hi-C protocols such as dilution Hi-C [2], *in situ* Hi-C [3] and DNase Hi-C [4]. We did not include data from other types of Hi-C-derived protocols such as Capture Hi-C [13] and HiChIP [19], as they are designed for detecting interactions enriched for specific elements such as promoters and transcription factors. We also excluded data from Methyl-HiC [20], which is useful for cell identity and DNA methylation profiling. In total, we established a list of 348 Hi-C datasets from human cell lines and tissues, generated from 30 individual studies (Supplementary Table 1).

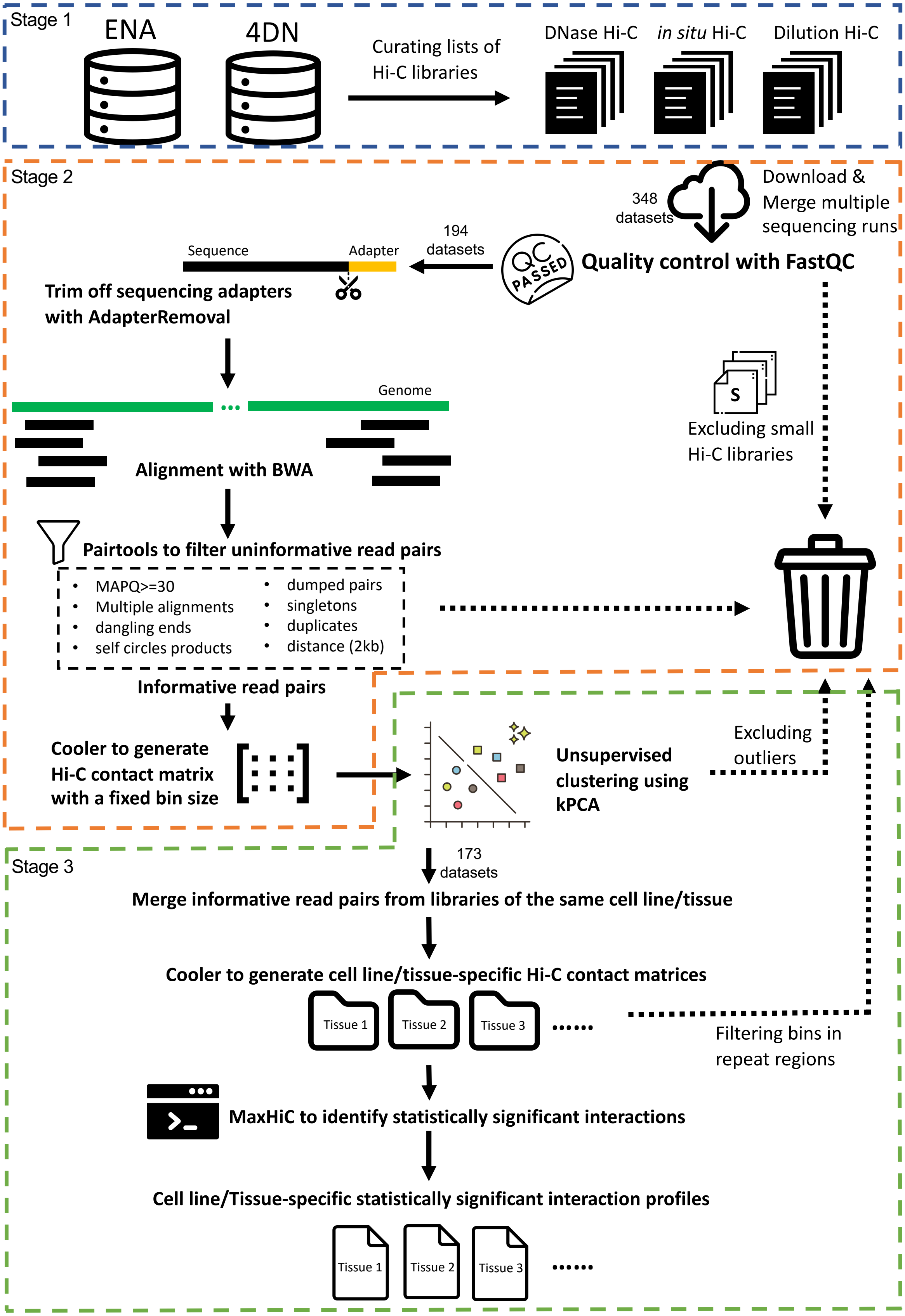

Stage 1
ENA
4DN
Curating lists of
Hi-C libraries
DNase Hi-C
in situ Hi-C
Dilution Hi-C
Stage 2
348
datasets
Download &
Merge multiple
sequencing runs
194
datasets
QC
PASSED
Quality control with FastQC
Sequence
Adapter
Trim off sequencing adapters
with AdapterRemoval
Genome
Alignment with BWA
S
Excluding small
Hi-C libraries
Pairtools to filter uninformative read pairs
• MAPQ>=30
• Multiple alignments
• dangling ends
• self circles products
• dumped pairs
• singletons
• duplicates
• distance (2kb)
Informative read pairs
Cooler to generate
Hi-C contact matrix
with a fixed bin size
Unsupervised
clustering using
kPCA
Excluding
outliers
Stage 3
173
datasets
Merge informative read pairs from libraries of the same cell line/tissue
Cooler to generate cell line/tissue-specific Hi-C contact matrices
Tissue 1
Tissue 2
Tissue 3
......
Filtering bins in
repeat regions
MaxHiC to identify statistically significant interactions
Cell line/Tissue-specific statistically significant interaction profiles
Tissue 1
Tissue 2
Tissue 3
......

Figure 1: A schematic view of the customised computational pipeline used in this study. Cell line/tissue-specific interaction profiles were taken from 348 public Hi-C datasets and filtered down to 173 profiles after quality control and filtering. In stage 1, data was collected from databases including ENA and 4DN data portal and then QC was performed (Stage 2) on raw sequencing data and filter datasets with bad quality, followed by conduct mapping, alignment filtering and generating Hi-C contact matrices. In stage 3, datasets were merged from the same cell lines and tissues, and statistically significant interaction profiles generated for each cell line/tissue using MaxHiC.

After data collection, we established a custom data processing pipeline based on best-practise Hi-C data processing protocols [18] to analyse all Hi-C datasets (Figure 1). We first conducted quality control on raw sequencing data of 348 Hi-C datasets (Supplementary Figure 1) and removed datasets based on strict selection criteria. Based on the aim of generating cell line/tissue-specific interaction profiles at high resolution (10 kb), we only kept datasets with more than 90 million reads and removed smaller datasets that would not provide sufficient resolution. Furthermore, we removed 5 libraries that were constructed without the HiC cross-linking step and served as a control to the protocol development [3]. These filtering steps resulted in 154 datasets being removed, leaving 194 for further processing (Stage 2, Figure 1) (Supplementary Table 2).

Of 194 Hi-C datasets, the mean and median of raw read counts were 345,200,859 and 227,641,938, respectively (Supplementary Table 3). The largest dataset is from a Brain tissue dataset (PRJNA661621) containing over 3 billion raw reads, and the smallest is a LNCaP cell line from study GSE73785 [21] containing just over 91 million reads. We observed an average mapping rate of 97.09% to the human hg38 genome [22], with on average 61.73% read pairs (of raw sequencing reads) having mapping quality over 30 (See Methods Chapter). Using Pairtools [23] to remove uninformative read pairs, including low-mapping quality pairs (MAPQ < 30), multiple alignment pairs, singletons, duplicates, self-ligation products and read pairs with short distance (< 2 kb), we identified informative interactions representing *bona fide* chromatin interactions (Figure 1). The average informative interaction rate was 51.98% (of raw reads) with an average intra-chromosomal interaction rate (interactions between regions within the same chromosome) of 37.09% and inter-chromosomal interaction rate (interactions between regions in different chromosomes) of 14.89% (Supplementary Table 3). Finally, we utilized 10 kb as a fixed bin size for all Hi-C libraries to generate Hi-C contact matrices using cooler [24].

Unsupervised clustering of Hi-C samples

The Hi-C contact matrices, containing information on the frequency of interacting 10 kb genomic bin regions, can vary across tissues/cell-lines due to technical biases including sequencing batch effects, different HiC protocols, sequencing strategies and platforms. Therefore, in order to reduce these effects and ensure the quality of the cell line/tissue-specific profiles, we performed an unsupervised clustering with the intra-chromosomal interactions of 194 Hi-C libraries to identify potential outliers for each cell line/tissue. While principal component analysis (PCA) are commonly used to cluster genomics data [25–27], kernel principal component analysis (kPCA) [28] were found to outperform PCA by its capability of exploring higher order information in the input data, we therefore decided to use kPCA to carry out the unsupervised clustering of chromatin interactions.

In order to reduce the burden on computational resources, we conducted kPCA on intra-chromosomal interactions for chromosomes 1, 12 and 22 of 10 random Hi-C libraries, observing a consistent pattern of clustering across different chromosomes (Supplementary Figure 2). From this we then chose to use a single chromosome to represent the clustering of the whole genome, and performed kPCA on the intra-chromosomal interactions of chromosome 22 to visualise the relationships among Hi-C datasets (Figure 2, Supplementary Figures 3 and 4). Initially, we found that the Hi-C datasets were mostly separated by the type of HiC library protocol used, regardless of cell line or tissue type is used (Figure 2A), with the second largest source of separation being their cell line/tissue types (Figure 2B, Supplementary Figures 3 and 4). More importantly, outlier datasets that failed to cluster with other datasets with the same cell lines or tissues could also be spotted in the kPCA plots (Figure 2). Outliers are likely to introduce biases when we try to merge the Hi-C datasets of the same cell lines and tissues into one profile, leading to inaccuracy identification of cell line/tissue-specific interactions. We observed outliers in a number of cell lines/tissues, including GM12878, HUVEC, Jurkat T lymphocytes, Liver, T cell, NHEK and PrEC (Supplementary Figures 3 and 4), and removed 10 outliers datasets (i.e. failed to clustered with others) from the subsequent analysis.

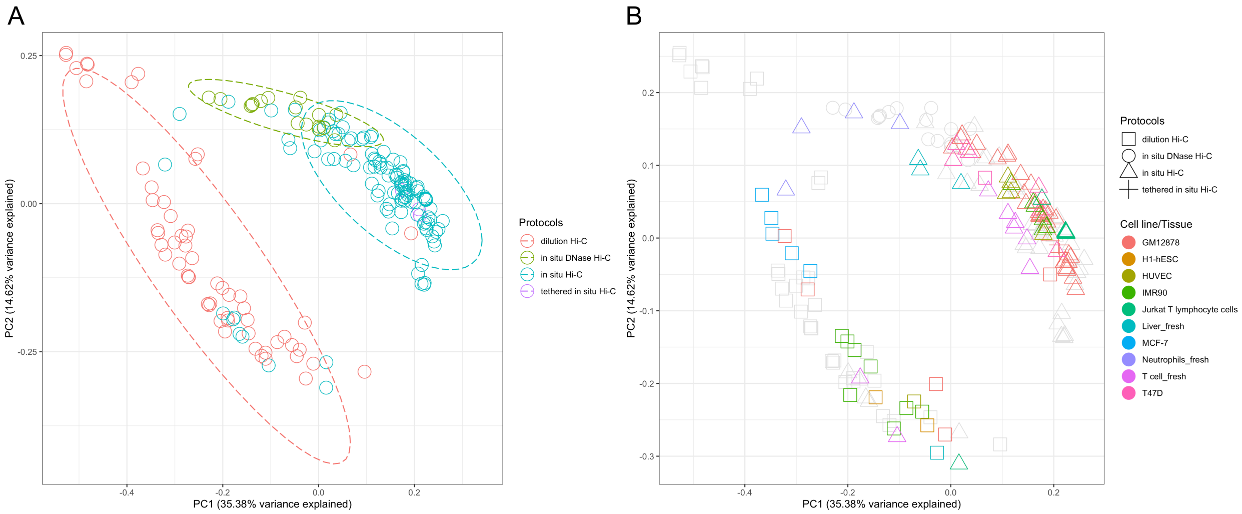


Figure 2: Kernel principal component analysis (kPCA) of intra-chromosomal interactions of chr22 of 196 Hi-C libraries, (A) coloured by Hi-C protocols and (B) the cell line or tissue sampled in more than one study. Libraries that are found in only one study are colored in grey.

Additionally, datasets from IMR90 (Supplementary Figure 3), H9-hESC and HMEC (Supplementary Figure 4) seem to form two separate clusters by protocol. This separation is likely due to the difference of the digestion enzyme, with 4-bases cutter MboI used in *in situ* Hi-C protocols and 6-bases cutters such as *HindIII* being used in dilution Hi-C, leading to different capturing efficiency of chromatin interactions. In order to generate Hi-C contact profiles with higher resolution, for these 3 cell lines, we removed the cluster of datasets generated by dilution Hi-C, with this filtering removing a total of 21 datasets, leaving 173 Hi-C datasets to be used to generate cell line/tissue-specific interaction profiles (Figure 1).

Generation of statistically significant interaction profiles

In order to generate MaxHiC-detected statistically significant interaction (MADSSI) profiles for each cell line and tissue we curated, we merged samples based on their cell line and tissue types to generate cell line/tissue-specific interaction pairs using 173 Hi-C datasets. Some interaction pairs are characterised by only 1 sample, such as adrenal and aorta tissues, while others were composed of more than 5 datasets, such as Jurkat T lymphocytes and skin fibroblast tissue (Supplementary Figure 5). Interaction pairs of the GM12878 cell line were merged from 36 datasets, much more than any other cell lines/tissues (Supplementary Figure 5). This sample merging step resulted in 51 informative interaction pairs files from individual cell lines and tissues (Supplementary Table 4).

Interaction pairs were then used to construct individual cell line/tissue-specific Hi-C contact matrices with a fixed bin size of 10 kb (Figure 1). It is necessary to perform bin-level filtering once the Hi-C contact matrices are generated, particularly removing genomic bins located in repeat regions where potential alignment issues are likely to occur, impacting the accuracy of results [29]. We therefore removed interaction bins located in repeat regions by removing bins that had over 50% overlap with annotated repeat regions and blacklist regions in the hg38 genome [30,31] (Supplementary Table 5). We then used hierarchical clustering with the WARD algorithm [32] to visualise the relationships between the Hi-C contact matrices of 51 cell lines/tissues (Figure 3). All 51 contact matrices clustered into 5 clusters, with the largest cluster (yellow/green) containing 19 cell lines and tissues. In this cluster, we found some cell lines and tissues that are biologically similar, such as left and right heart ventricle tissues as well as hippocampus and cortex. Similar cases are observed in other clusters as well, including LNCaP and PC3 cell lines (purple cluster, both are prostate cancer cell lines), 786-M1A and 786-O_TGL cell lines (light blue cluster, both are renal cancer cell lines), HSPC and B cells (red cluster), and Jurkat cell lines and T cells (green cluster) (Figure 3). However, some exceptional cases of biologically similar cell types that don't cluster together are observed, such as B cells and GM12878 are assigned to different clusters.

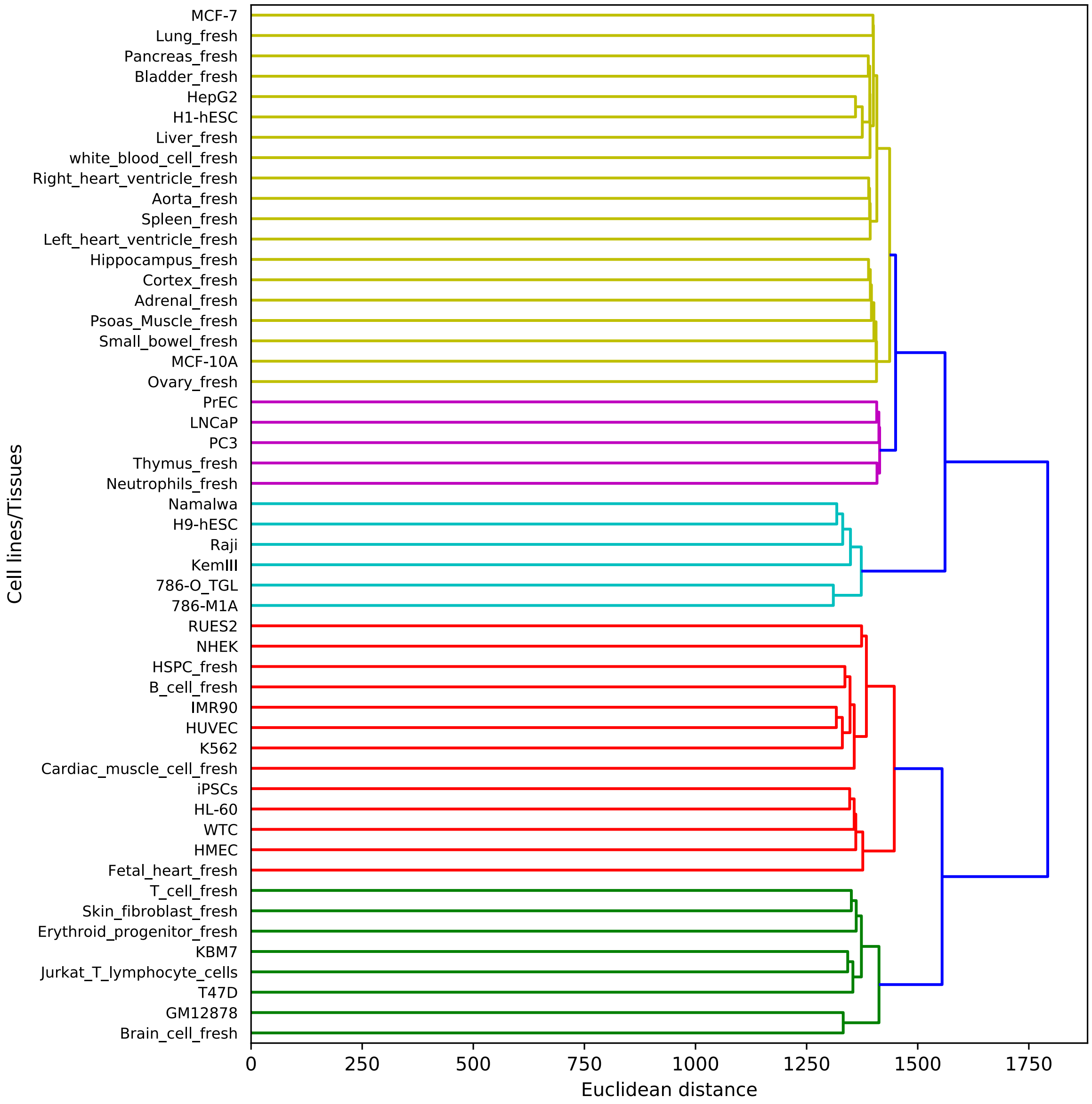


Figure 3: Hierarchical clustering of cell line/tissue-specific Hi-C contact map showing relationships between cell line/tissue using the WARD algorithm. The X axis is the euclidean distance between the contact maps of each group, and the Y axis are contact maps of 51 cell lines and tissues. Different colours of the branches indicate different clusters detected by the algorithm.

After extensive quality control and filtering, we then aimed to identify statistically significant chromatin interactions using MaxHiC. We extracted intra-chromosomal statistically significant interactions (referred from now on a standard "interactions") for 51 cell lines/tissues using an adjusted p-value threshold of 0.05 (Benjamini Hochberg correction) to generate MaxHiC-detected statistically significant interaction (MADSSI) profiles (Supplementary File 1). Although inter-chromosomal interactions have been shown to exist and may play a role in promoting chromatin structure formation [33–35], we instead chose to only analyse intra-chromosomal interactions which are more likely to impact gene expression and regulation [34]. Of 51 MADSSI profiles, the average interaction count was 125,865, with cell-line GM12878 cell line containing the most MADSSI (1,614,240), while Adrenal tissue contained the least containing only 6 MADSSI (Figure 4 and Supplementary Table 6). Despite the differences in GM12878 and Adrenal tissue, the mean distributions of sequencing read pair count per interaction were similar across all 51 cell lines/tissues (Figure 4).

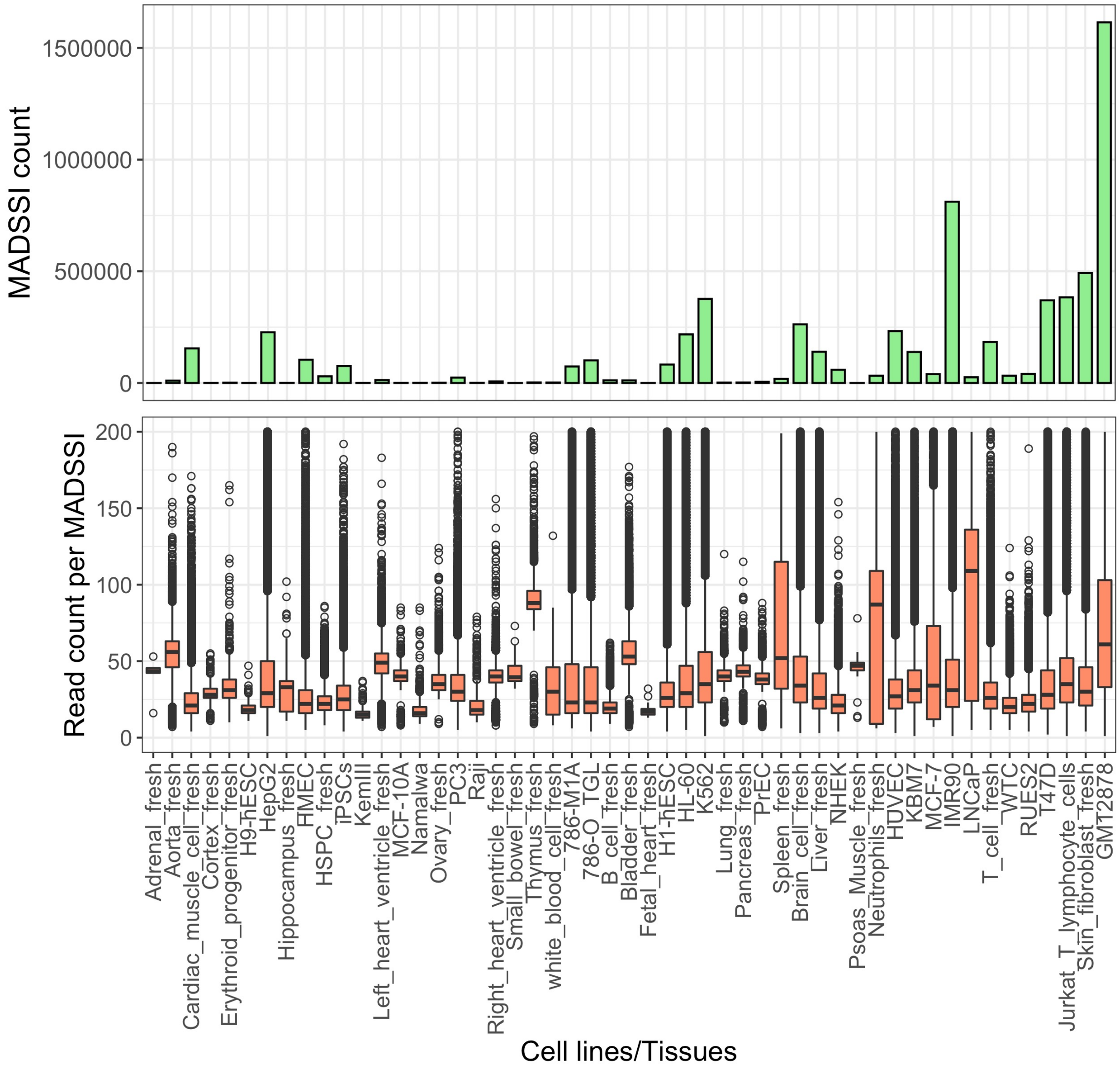
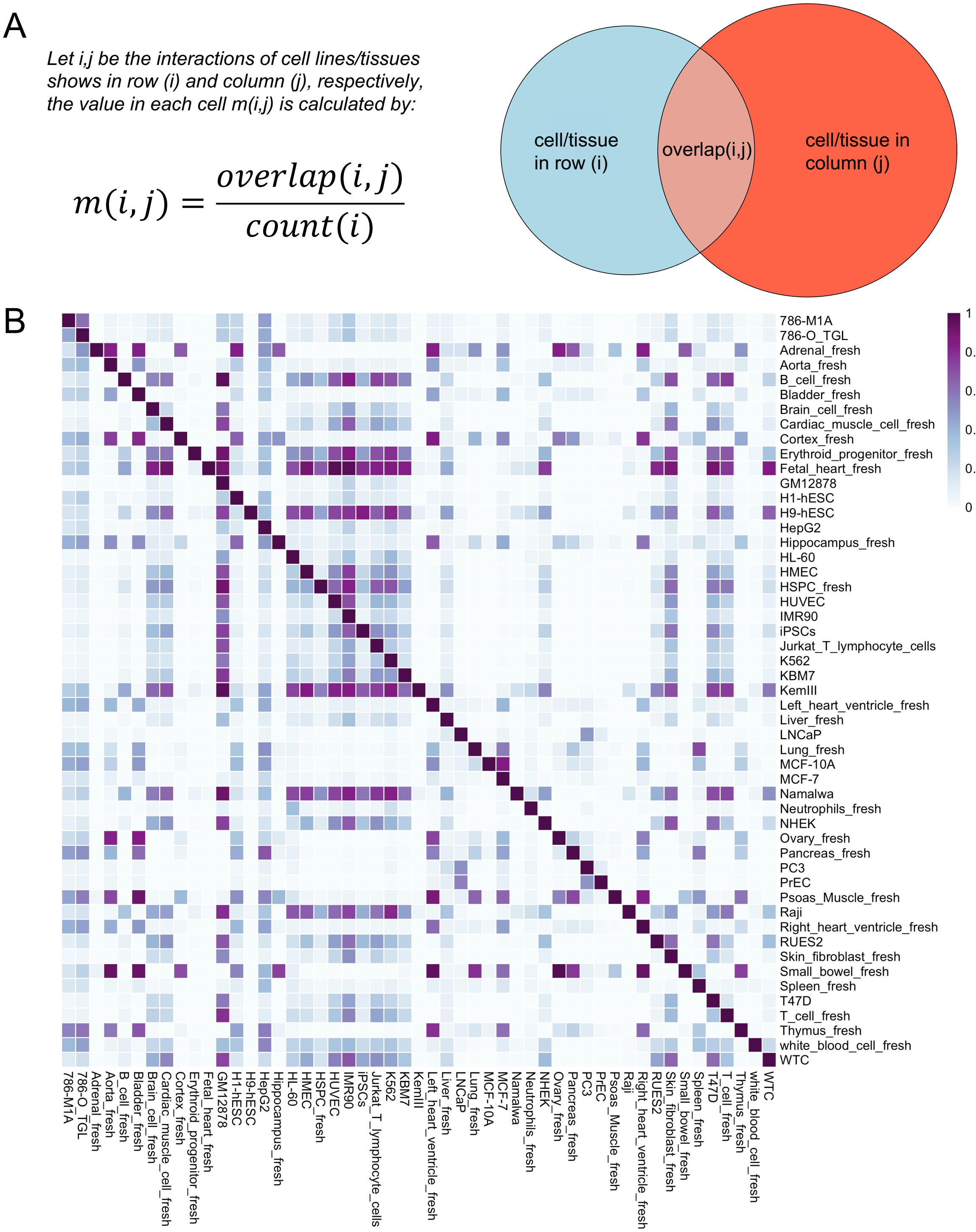


Figure 4: Output summary statistics from designed HiC workflow in this chapter. (Top panel) Intra-chromosomal statistically significant interaction count detected by MaxHiC across 51 selected cell lines/tissues. (Bottom panel) Distribution of read pair count per intra-chromosomal statistically significant interactions in 51 selected cell lines/tissues.

Unique and non-unique statistically significant interactions

After generating MADSSI profiles across our 51 cell-line/tissue sets, we then described the uniqueness of each cell/tissue-type and whether common statistically significant interactions exist in all datasets. Comparing MADSSI profiles between each cell line and tissue by calculated the percentage of interaction in one cell line and tissue that are also found in others (Figure 5), we observed that interactions of some cell lines/tissues are not uniquely identified in their corresponding cell lines/tissues, but found in other cell lines/tissues, such as interactions detected in B cells, erythroid progenitor cells, fetal heart, H9-hESC, HSPC, KemIII, Namalwa, NHEK, psoas muscle, small bowel and WTC. These cell lines and tissues have more than 37.39% (on average) of MADSSI also identified in others (Figure 5B). Furthermore, we found that more than 50% MADSSI of all analysed cell lines and tissues are MADSSI of GM12878 and IMR90 cell lines. This is likely due to GM12878 and IMR90 containing the most number of interactions in their MADSSI profiles, indicating that sampling size has an impact detecting uniqueness across samples (Figure 4).



Let i,j be the interactions of cell lines/tissues shows in row (i) and column (j), respectively, the value in each cell m(i,j) is calculated by:

$$m(i,j) = \frac{overlap(i,j)}{count(i)}$$

Figure 5: Analysis of comparing MADSSI among 51 cell lines and tissues. (A) an equation (left) used to calculate values representing the non-uniqueness of interactions in each cell/tissue compared to others; and a venn diagram (right) demonstrating the elements in the equation. (B) Heatmap shows the fraction of MADSSI in each cell line/tissue that are also identified in other cell lines/tissues. The value in each cell is calculated by the equation in A.

Frequency of MADSSI found in cell lines and tissues were then used to define unique (i.e. detected in one) and non-unique interactions (detected in more than one cell lines/tissues). Of all MADSSI, 62.3% of the interactions are found in only one cell line/tissue, defined as unique interactions, while the rest of MADSSI (38%) are defined as non-unique interactions (Supplementary Figure 6), suggested that tissue-specific interactions are the dominant form of significant interactions. In non-unique interactions, the most frequent MADSSI is the gene body of *BCL6* (chr3:187,730,000-187,740,000) contacts a region (chr3:188,940,000-188,950,000) containing *TPRG1-AS1* and *TPRG1* gene. This MADSSI is observed across 35 cell lines and tissues, such as T cells, thymus, spleen, K562 and HVEC etc, however from our knowledge it has never been investigated in any specific study.

We investigated the distance between interacting anchor bins of unique interactions and non-unique interactions. Unique interactions generally occurred at longer linear distances compared to non-unique interactions, with the mean unique genomic distance being 1.9 Mb (median = 650 kb) compared to 616 kb (median = 400 kb). The longest unique interaction was found in GM12878, spanning 227.8 Mb on chromosome 1 and contacting *CAMTA1* gene and a non-coding region, while the size of chromosome 1 is over 248.9 Mb. And the longest non-unique interaction was also detected in chromosome 1 in GM12878 and brain cells, spanning 219.15 Mb and contacting *DDI2* gene and a non-coding region. Interestingly, these two interactions both contact a non-coding region on chromosome 1 (chr1.q42.3) (Supplementary Figure 7), where layered H3K27ac histone markers and distal enhancer signatures are found, suggesting a conical enhancer hub. Comparing the distance distribution of unique and non-unique interactions, more proportions of non-unique interactions are observed in short distances. For example, 38.61% of non-unique interactions occur at distances less than 300 kb, compared to 26.91% of unique (Figure 6). However at long distances, such as 700 kb, unique interactions become the dominant form of interactions (Figure 6).

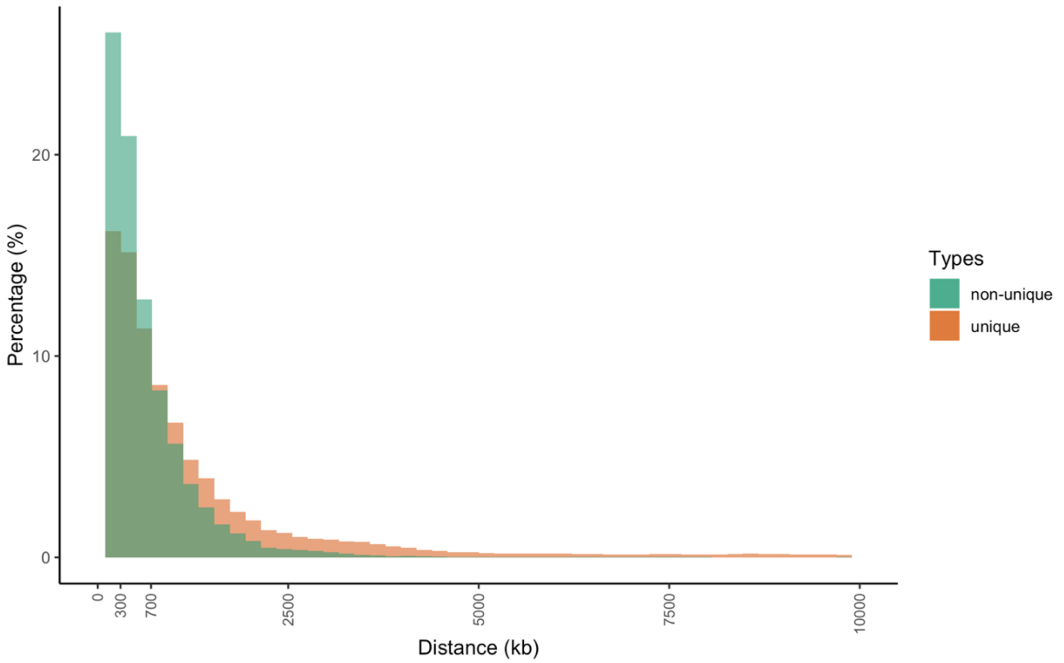


Figure 6: Distance distribution of unique and non-unique MADSSI across 51 cell lines/tissues. In order to fairly compare non-unique MADSSI to unique MADSSI and account for the different interaction count of them, we used percentage of interaction in the y axis.

If unique MADSSI were only found in one tissue, we then investigated whether specific types of genes were contacted by carrying out Gene ontology (GO) enrichment analysis to detect enriched GO terms that are associated with genes found within these regions of unique interactions (Supplementary Table 9). We found out that the unique interactions were enriched for GO terms that correspond to the specific cell types or tissues. For example, “T cell differentiation” (GO:0030217) and “T cell activation” (GO:0042110) are uniquely found enriched for genes contacted by T cell unique interactions, “Fc receptor signaling pathway” (GO:0038093) for B cell, “ERBB2 signaling pathway” (GO:0038128) for MCF-7 cell line and “bundle of His cell to Purkinje myocyte communication” (GO:0086069) for right heart ventricle (Figure 7).

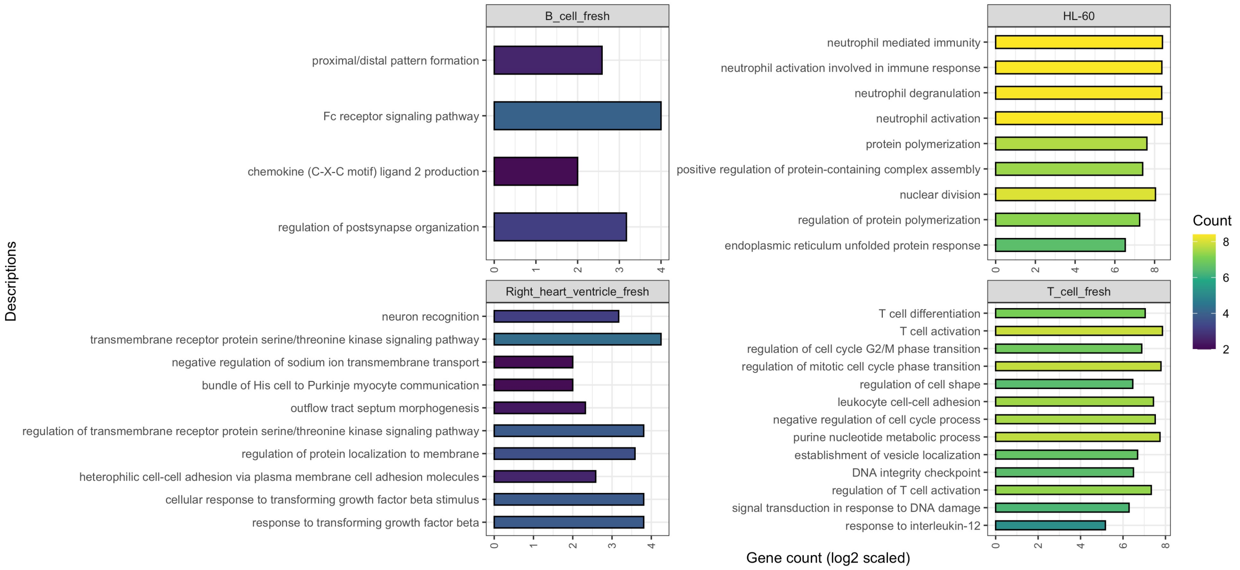


Figure 7: Gene ontology (GO) over-representation analysis of genes contacted by uncommon interactions (statistically significant interactions observed in only one or two cell lines/tissues) across 52 cell lines/tissues. For better visualisation, unique enriched GO terms of four cell lines/tissues are selected here. Full lists of enriched GO terms across 52 cell lines/tissues can be found in Supplementary Table 9.

To demonstrate a specific example of unique MADSSI profiles interacting with target genes, we found that *SATB1*, a gene that regulates the action of *FOXP3* in regulatory T cells via 3D chromosome structure [36–40], is contacted by MADSSI uniquely found in T cell samples (Figure 8). We also observed T cell-specific CTCF-binding sites and interactions between the promoter of *SATB1* and multiple non-coding regions with active enhancer chromHMM states upstream, demonstrated by the green coloured states on the bottom genome browser panel of Figure 9, indicating potential regulation occurring via the formation of CTCF-derived promoter-enhancer loop. Altogether, this example demonstrates cell/tissue type specificity of unique interactions and these interactions might govern the regulations of cell/tissue type-specific pathways that are important for key transcriptional regulators.

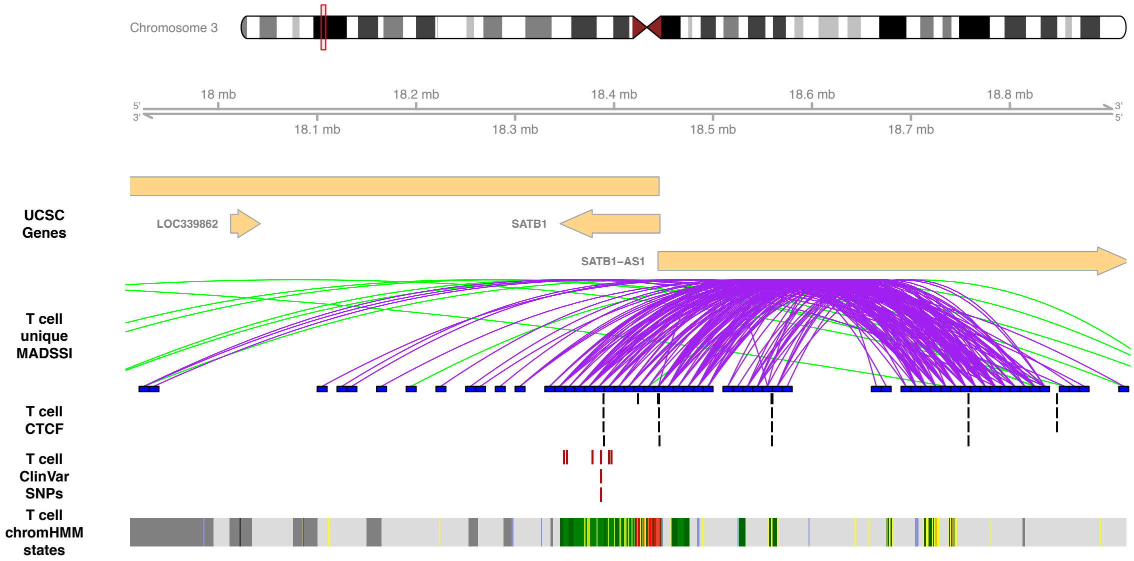


Figure 8: An example of unique MADSSI contacting cell line/tissue-specific genes across the SATB1 gene loci of chromosome 3. The genomic tracks plot shows the T cell-unique MADSSI (purple coloured are interactions within the plotting area and the green ones are interactions linking regions outside) around the *SATB1* region, along with T cell-specific CTCF-binding sites and T cell chromHMM states (color legend provide in the top left corner), which are obtained from the roadmap epigenomics project [41].

Given their propensity to impact multiple tissues rather than one, non-unqiue interactions may also play a more canonical role in gene regulation. We therefore ranked non-unique interactions by their membership across cell-lines/tissues, and focused on the top 10 non-unique interactions (Supplementary Table 7). These interactions overlapped with 20 genes: 18 protein-coding genes, one long non-coding RNA (LOC100131635) and one micro-RNA (MIR4315-1) (Supplementary Table 8). Many of these genes have shown to be highly associated with disease, such as *RNASEL* was demonstrated to be a candidate hereditary prostate cancer gene [42,43], SRSF1 and MRPS23 were proposed to contribute to breast cancer [44,45] and NYNRIN was shown to be a predisposition gene for Wilms tumour [46].

Based on the gene expression data from the EBI-Expression Atlas [47], 16 of the 18 protein-coding genes exhibit relatively high expression levels across a wide range of tissues and cell types, the exceptions being gene *LRRC52-AS1* and *TPRG1-AS1* (Figure 9). *TPRG1-AS1* is mildly expressed in liver, breast, esophagus, adrenal gland and adipose tissue, while *LRRC52-AS1* is not expressed in any of the chosen cells and tissues (Figure 9). Additionally, the promoter regions, defined here as the region 2 kb upstream of the transcription start site, of most of these genes overlap with H3K27ac ChIP-seq peaks in a number of cell lines and tissues, except for gene *LRRC52-AS1* (Figure 9). Together these indicate that the top 10 non-unique interactions that are observed in multiple cell lines and tissues may play important roles in the regulation of these commonly expressed genes.

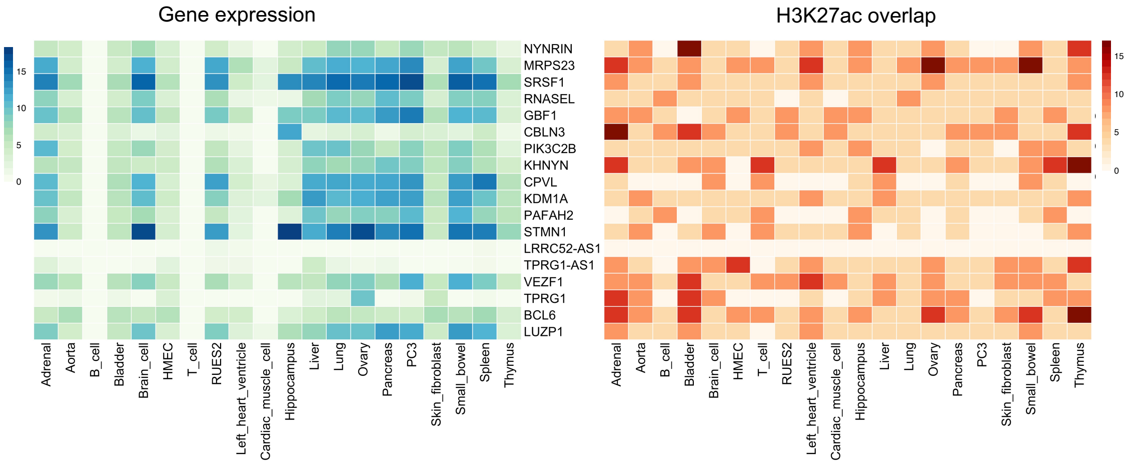


Figure 9: Normalised (log2-scaled) gene expression profile of genes contacted by top 10 non-unique interactions in different cells and tissues (left), obtained from the EMBL-EBI Expression Atlas [47]. Additionally, promoters of the same genes were compared to H3K27ac ChIP-seq peak information (right) obtained from the ENCODE project [48].

Accumulated interactions form “hot zones”

So far, we have identified that unique, tissue-specific MADSSI are more common than non-unique interactions across multiple tissues/cell-types, making up over 62.3% of interactions. We next focused on whether MADSSI common to multiple tissue/cell-lines were co-located in specific areas, indicative of a spatial regulatory mechanism. Given the structural importance of chromatin interactions, we hypothesised that an accumulation of common interactions (i.e. found in many tissues/cell-lines) in specific bins or genomic regions will identify features that are essential for the structural integrity of the chromosome arrangement, features such as insulator transcription factor CTCF-binding sites and TAD boundaries. In order to reveal genomics regions with statistically significant interaction across many cell lines/tissues, we took 10 kb genomic bins from 51 cell line/tissue-specific MADSSI profiles and ranked each bin by their presence in the number of cell lines/tissues. These genomic bins, which we name “hot zones”, are defined as interacting regions observed with more than half of all cell lines/tissues (at least 26 cell lines/tissues).

Using this new metric, we identified a total of 2,442 interaction hot zones dispersed across the genome (Supplementary Table 10). Hot zones can be easily identified on the Hi-C contact maps when visualising them with MADSSI profiles (Figure 10A). For example, the interaction hot zone chr3:187730000-187740000 (Figure 10), which is the most frequently shared interaction bin (found in 38 cell lines/tissues), was connected by a large number of MADSSI to many other genomic bins. This specific hot zone was first observed in the most frequent non-unique interactions described previously, within the gene *BCL6*, which is a known transcription factor that plays an important role in the differentiation of T cells (Figure 10B) [49]. Interestingly, we can also see hot zone patterning in samples/tissues with poor sequencing coverage, such as MCF-7 featured in Figure 10, suggesting that both MADSSI and hot zone identification can be used to outline structurally important regions in spite of technical limitations.

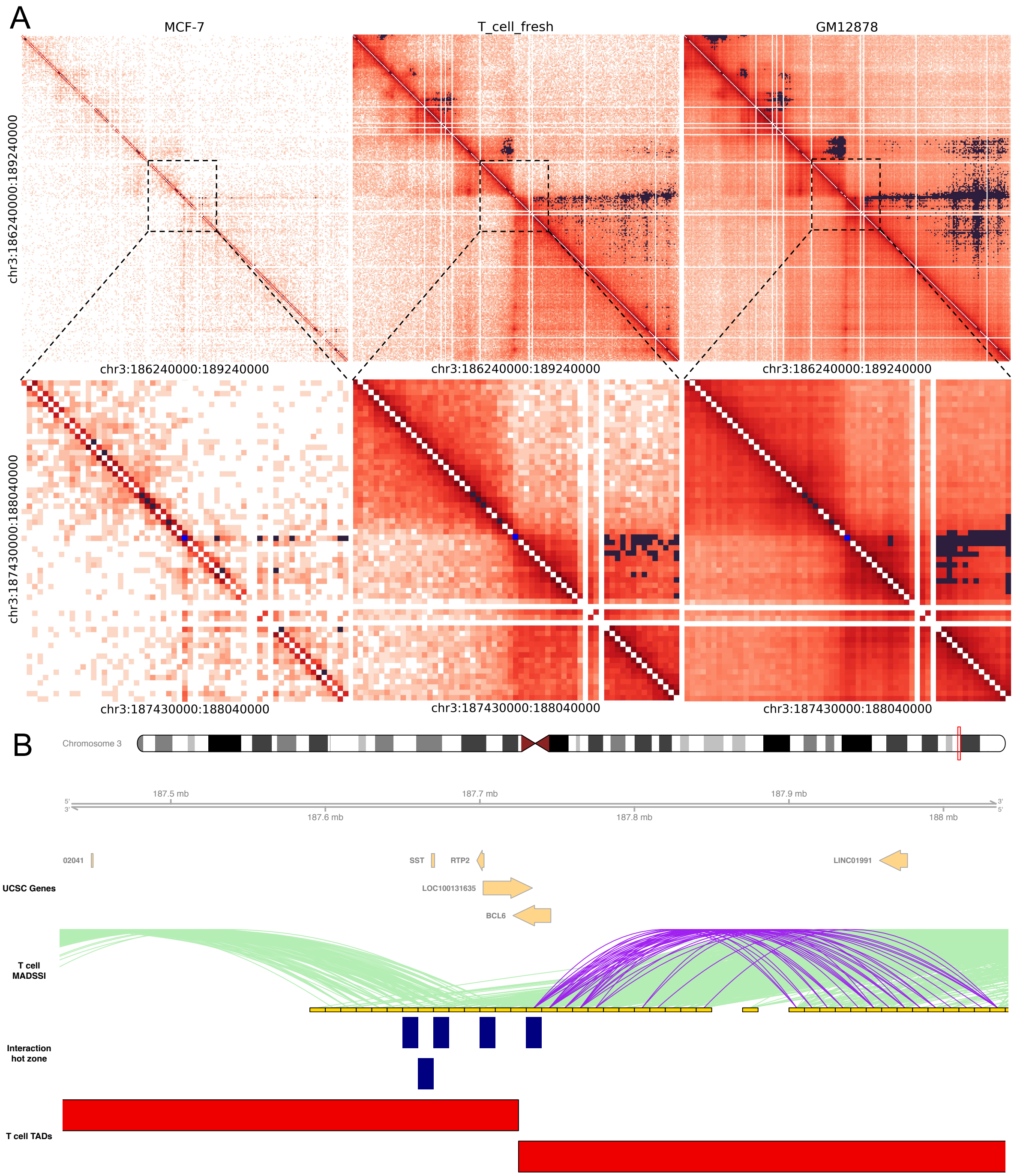


Figure 10: Visualisation of MADSSI and interaction hot zones across the gene loci BCL6. (A) Hi-C contact map of MCF-7, T cell and GM12878. Top panel: a 3 Mb region in chromosome 3, black dots in the top right triangle indicate MADSSI identified in the corresponding cell line/tissue, black dots in the diagonal indicate interaction hot zones identified in this study. Bottom panel: a 60 kb region zoom from the top panel as marked. The blue dot in the diagonal indicates the interaction hot zone chr3:187730000-187740000. (B) genomic tracks plot shows interaction hot zones, T cell-specific MADSSI and TADs with gene annotations in the same region as the bottom panel of (A).

We also observed that the hot zones tend to be located close to the boundary of the interacting domain shown on the heatmap (Figure 10A). We therefore identified TADs across all 51 cells/tissues (Supplementary File 2) to investigate the relative distance between interacting hot zones and TAD boundaries. Consistent with the observation of finding hot zones near the boundary of interacting domains (Figure 10A), we found that the majority of interacting hot zones were located close to the TAD boundaries, with an average of 74.36% located less than 125 kb to TAD boundaries given a mean TAD size of 576 kb (Supplementary Figure 8), confirming our hypothesis of the structural nature of these regions and suggesting the hot zones may serve an important role in maintaining the structural integrity of chromosomes. Compared to other regions, interaction hot zones are also significantly enriched for H3K27ac and H3K4me1 histone markers and CTCF-binding sites in cell lines/tissues where such annotation data are available (Figure 11A). Hot zones that are observed in T cells for example overlap 50.3% of H3K4me1 histone peaks annotated in T cells (Fisher’s exact test p-value = 1.39e-05) (Supplementary Figure 9). We also found hot zones have approximately 3-fold enrichment for CTCF-binding sites compared to other regions, and 2-fold enrichments were observed for H3K27ac and H3K4me1 histone markers (Figure 11A), indicating that hot zones structurally defined by CTCF could play a role in maintaining canonical gene expression and regulation.

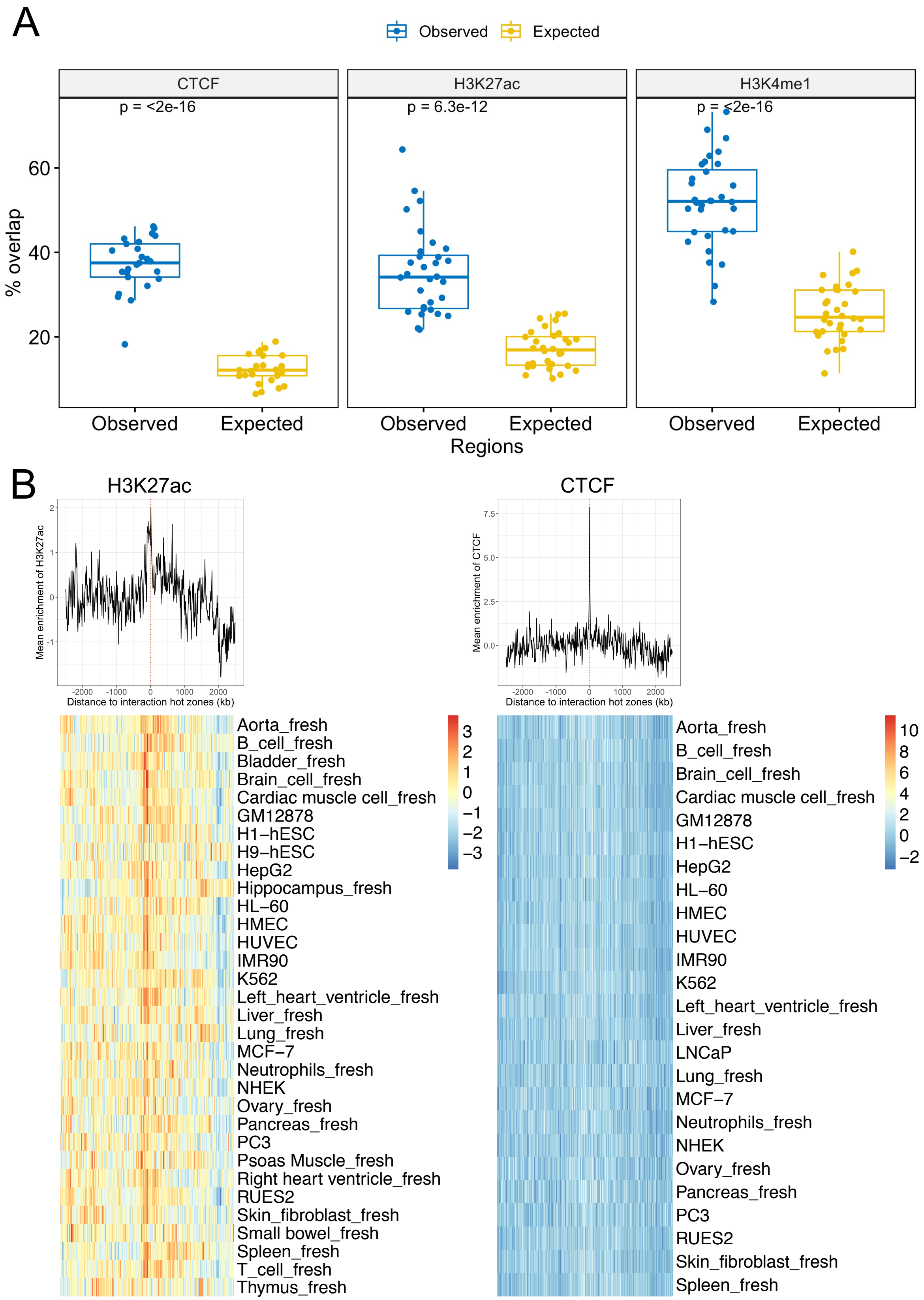

A
Observed
Expected
CTCF
H3K27ac
H3K4me1
p = <2e-16
p = 6.3e-12
p = <2e-16
% overlap
60
40
20
Observed
Expected
Observed
Expected
Observed
Expected
Regions
B
H3K27ac
Mean enrichment of H3K27ac
2
1
0
-1
-2000
-1000
0
1000
2000
Distance to interaction hot zones (kb)
CTCF
Mean enrichment of CTCF
7.5
5.0
2.5
0.0
-2000
-1000
0
1000
2000
Distance to interaction hot zones (kb)
Aorta_fresh
B_cell_fresh
Bladder_fresh
Brain_cell_fresh
Cardiac muscle cell_fresh
GM12878
H1−hESC
H9−hESC
HepG2
Hippocampus_fresh
HL−60
HMEC
HUVEC
IMR90
K562
Left_heart_ventricle_fresh
Liver_fresh
Lung_fresh
MCF−7
Neutrophils_fresh
NHEK
Ovary_fresh
Pancreas_fresh
PC3
Psoas Muscle_fresh
Right heart ventricle_fresh
RUES2
Skin_fibroblast_fresh
Small bowel_fresh
Spleen_fresh
T_cell_fresh
Thymus_fresh
3
2
1
0
−1
−2
−3
Aorta_fresh
B_cell_fresh
Brain_cell_fresh
Cardiac muscle cell_fresh
GM12878
H1−hESC
HepG2
HL−60
HMEC
HUVEC
IMR90
K562
Left_heart_ventricle_fresh
Liver_fresh
LNCaP
Lung_fresh
MCF−7
Neutrophils_fresh
NHEK
Ovary_fresh
Pancreas_fresh
PC3
RUES2
Skin_fibroblast_fresh
Spleen_fresh
10
8
6
4
2
0
−2

Figure 11: Enrichment analysis with interaction hot zones. (A) global enrichment analysis shows the enrichment of CTCF-binding sites, H3K27ac and H3K4me1 histone marks in hot zones. The expected overlaps are calculated the overlaps between permuted 10 kb regions across the genome and annotations while the observed overlaps are between hot zones and annotations. The p-values are calculated using a t test. (B) local enrichment analysis shows the enrichment of H3K27ac marks and CTCF-binding sites on hot zones compared to the surrounding regions. Top panel: mean of enrichment score of a 5 Mb region centred at interaction hot zones across cell lines/tissues. Bottom panel: enrichment score of a 5 Mb region centred at interaction hot zones in each cell line/tissue.

We then used local enrichment tests to characterise the histone modifications, CTCF, and candidate cis-regulatory elements (cCREs) signatures [50] across 5 Mb regions centered at interaction hot zones. Interaction hot zones are locally enriched for H3K27ac and H3K4me1 histone markers, which are often associated with active enhancers [51,52], in most of the cell lines/tissues (Figure 11B and Supplementary Figure 10), while no obvious patterns were found for H3K27me3 marker (Supplementary Figure 10), which is associated with downregulation of genes [53].

We also found a significant enrichment of CTCF-binding sites associated with hot zones across all cell lines/tissues (Figure 11B), consistent with the previous demonstration of hot zones being located close to TADs boundaries and globally enriched with CTCF-binding sites. With the strong CTCF-binding, interaction hot zones may be located near the base of chromatin loops, which are formed by the assistance of CTCF-binding [54]. Finally, we also used three signatures from cCREs to perform local enrichment, including distal enhancers (enhD), enhancer signals located more than 2 kb from transcription start site (TSS), proximal enhancers (enhP), enhancer signals located within 2kb from TSS, and promoters (Prom), regions with high DNase and H3K4me3 signals and located within 200 bp from annotated TSS [50]. We observed that the interaction hot zones are locally enriched for enhD and enhP (Supplementary Figure 10), but no obvious enrichment found with Prom. Consistently as we showed hot zones are most enriched for CTCF-binding sites globally, the local enrichment level of CTCF-binding sites are higher than other annotations (Supplementary Figure 10).

While the accumulation of statistically significant interactions has the potential to highlight structural interactions, other frequency interacting areas have been defined in previous studies, namely frequency interacting regions or FIREs that have shown to be regulatory relevant [55]. FIREs will be different to hot zones given they are generally called using one sample group, but we wanted to compare both to highlight potential regulatory mechanisms highlighted in Schmitt et al., 2016. Taking one cell-line sampled across both studies, i.e. GM12878, we compared 2,441 hot zones with 19,303 published GM12878-specific FIREs at 5 kb resolution, identifying 331 (13.56%) hot zones overlapping with 393 (2.04%) FIREs. In order to examine the potential structural and regulatory functionality of hot zones compared to FIREs, we compared hot zones and FIREs regarding their overlapping with GM12878-specific CTCF-binding sites, cCREs enhancer signatures (including enhP and enhD) and GM12878-specific H3K27ac marker (Figure 12). We found that hot zones are significantly more enriched for CTCF-binding sites and H3K27ac marker than FIREs (Fisher exact test, p-value < 2.2e-16), while similar large proportions of FIREs and hot zones overlapped cCRE enhancer signatures (Figure 12). This suggests that compared to FIREs, hot zones are more likely to be associated with structural integrity and may serve as a structural markers, and more indicative of structural function than other statistically-derived domains such as sub-TADs [56–58].

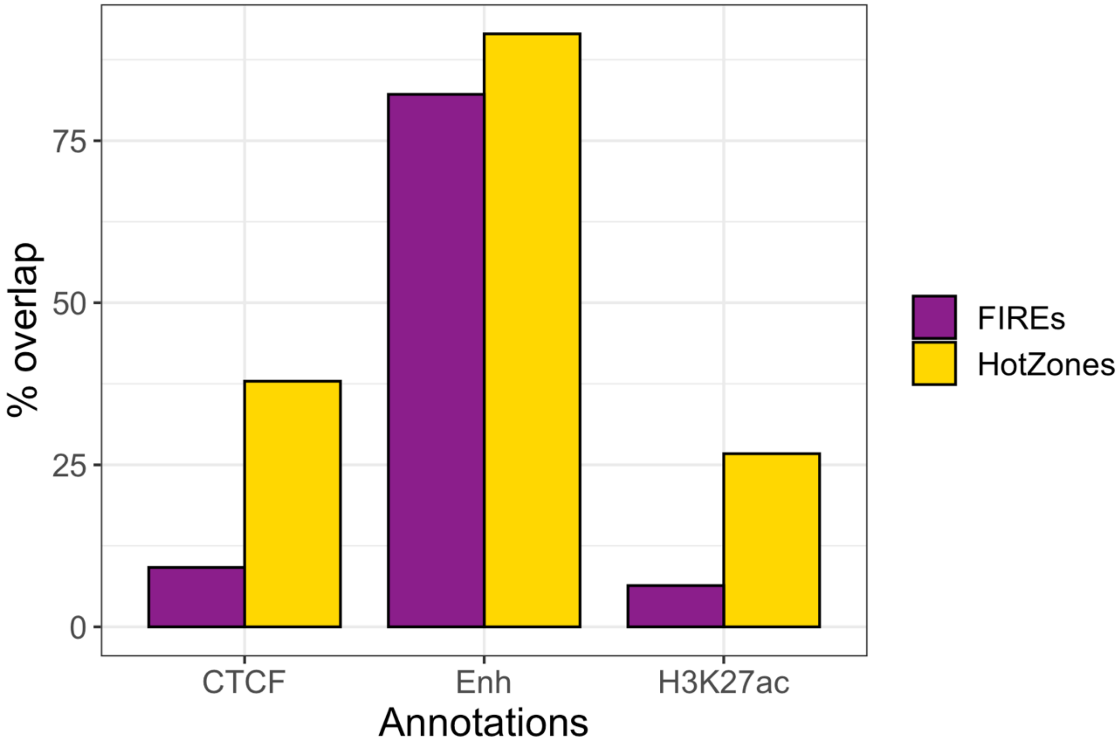


Figure 12: Comparison between GM12878 hot zones and FIREs of CTCF-binding sites and cCRE enhancer signatures overlapping status. GM12878 FIREs are obtained from a published study [55].

Discussion

Hi-C assays were developed to facilitate the detection of 3D chromatin interactions and the construction of 3D contact maps [2–4]. However, Hi-C sequencing datasets are complicated by various sources of biases such as random ligation artifacts and mappability of the DNA sequences [6]. Therefore, identifying interactions that are potentially more biologically functional than others is essential to future functional genomics studies. In a previous study, we showed that MaxHiC is able to pinpoint statistically significant interactions which show higher levels of overlap with regulatory annotations, such as active histone markers and CTCF-binding sites, than those detected by other methods [9]. In this study, we expanded the use of this algorithm, using a custom HiC-meta-analysis workflow and MaxHiC to generate statistically significant interaction maps across 51 human cell lines and tissues (Figure 1).

Despite similarities between tissues/cell-lines, hierarchical clustering of chromatin contact maps showed only minor levels of similarity based on the biological groups. Using the WARD method, the contact maps of some cell types that are biologically more relevant were found to be clustered together (Figure 3), such as left and right heart ventricle tissues and Jurkat cell line and T cell. However surprisingly, HL-60, which was developed as a model to study neutrophils [59,60], failed to cluster with neutrophils. Another similar case is that the contact map of B cells didn't cluster with GM12878 (Figure 4), a B-Lymphocyte cell-line. This may suggest that the 3D genome profile of cell lines may evolve significantly from their primary cell types after generations of passages, similar to the previous findings that high-passage cell lines exhibit different expression profiles compared to the primary cell types [61–63]. Therefore, such an effect may need to be taken into account when comparing Hi-C data of cell lines to investigate 3D regulations related to diseases, especially in light of accumulated interactions and hot zones identified in this study.

The high number of unique compared to non-unique interactions shown in this study implies that the cell line/tissue-unique interactions are highly involved in regulations that are cell/tissue-type specific. This is consistent with previous findings that promoter-centred interactions are highly cell type-specific [14] and that tissue-specific chromatin interactions tend to involve active regulatory elements [55]. Studies associating diseases and the affected tissues [64,65] showed the Mendelian diseases to be highly tissue specific. GWAS variants of complex diseases, such as autoimmune, neurodegenerative and cardiovascular diseases [66–69] were shown to have a large tissue-specific contribution, such as a recent study used pancreatic islet-specific chromatin interactions to reveal the candidate enhancers and risk loci for type 2 diabetes [15]. Therefore, the cell lines/tissues-specific MADSSI we have catalogued in this study can provide an extra layer of information when associating genetic variation to the regulatory mechanism that is dysregulated in disease systems.

Interaction hot zones are significantly enriched for both structural and regulatory elements, suggesting that they play a major role in cellular integrity and canonical function. By considering each interaction bin separately, rather than the link between two interacting bins [55], we defined interaction hot zones, regions with statistically significant interactions in over half of our catalogued cell lines/tissues. Hot zones are more likely to possess structural information (CTCF-binding sites) than regulatory information (enhancer signatures) than FIREs, indicating the important role that hot zones play in maintaining the structure integrity of the genome. While TADs and sub-TADs have recently been questioned for their identification being lacking in biological perspective and their low concordance among identification methods [56,57,70], the interaction hot zones can be a more suitable candidate to infer large structural arrangement in the genome, which is steady across many cell and tissue types.

Previous studies have shown that structural domains, such as TADs and A/B compartments, are not only consistent across human tissues, but also invariant across species groups [55,71]. Importantly, due to the invariant nature of TAD boundaries and CTCF-binding, the removal of demarcating regions can be a harbinger for disease impact [72–74]. By accumulating interactions, hot zones may provide additional disease marker information, potentially via Machine Learning approaches that can be trained to identify specific hot zone-like signatures using large public Hi-C datasets. To ultimately prove the potential of hot zones in this space, CRISPR knockout experiments, such as ones which have been used to remove TAD boundaries [72], could be used to prove their disease identification potential.

In conclusion, by re-analysing published Hi-C datasets and generating the first statistically significant interaction profiles of 51 human cell lines and tissues, we have catalogued interactions uniquely detected in each cell line/tissue and the ones that are shared across multiple cell lines/tissues, and revealed a set of interaction hot zones across many cell and tissue types. These results are a valuable resource to allow further and more specific investigation into the regulatory circuits governed by cell/tissue-specific 3D genome structures.

Methods

Code availability

The code for the customised Hi-C data processing pipeline used in this study and described in the results is available in GitHub: https://github.com/ningbioinfostruggling/CustomHi-CPipeline.

Data sources

Public Hi-C datasets were downloaded using Aspera [75]. The datasets used in this study are obtained from either the ENA database or the 4DN data portal, more comprehensive information including download links is documented in Supplementary Table 1.

Cell line/tissue-specific annotations

The gene expression information used in this study was obtained from the gene expression atlas database [47]. The histone and CTCF ChIP-seq data were obtained from the ENCODE project [48].

Data processing

FastQC [76] and ngsReports [77] were used to carry out quality control of the raw sequencing data. AdapterRemoval [78] was used to trim off sequencing adapters. BWA-mem [79] was used to conduct mapping, followed by using Pairtools [23] to perform filtering and deduplication of interaction. Cooler [24] was used to generate Hi-C contact matrix and MaxHiC [9] was finally used to identify statistically significant interactions.

Hierarchical and unsupervised clustering analysis

We conducted unsupervised clustering on Hi-C contact matrices using Kernel Principal Component Analysis (kPCA). The interactions of one chromosome (chromosome 22) of all samples were used to generate a binary matrix by inspecting the appearance of interactions in Hi-C datasets, i.e. if the interaction is observed in such cell line/tissue (marked as 1) or not (marked as 0). We then used the *KernelPCA* function from the *sklearn* library [80] with cosine distance kernel. Finally. PC1 and PC2 are used for plotting the unsupervised clustering. Hierarchical clustering was performed by obtaining the first 1 million interactions of each contact matrix. Contact matrices were first normalised using *ICE* to account for biases such as library size, enzyme cutting bias and ligation bias. The normalised matrices were then used to calculate the linkage between each cell line/tissue based on the Ward's method [32] of Euclidean distance using the *cluster.hierarchy* function from the *scipy* library [81].

Identification of topologically-associated domains

To identify topologically-associated domains (TADs) for 51 cell lines/tissues, Hi-C contact matrix of 25 kb bin size of each cell line/tissue generated by cooler [24] were used to transform into dense matrix using the *sparseToDense.py* script from HiC-Pro [82]. Then the dense matrices were used by TopDom [83] to identify TADs.

Global and local enrichment analysis

For global enrichment analysis, we calculated the percentage of interaction hot zones that overlapped with the corresponding annotations in each cell line/tissue with annotation data available, then comparing it to the expected percentages of overlapping, which calculated by permuting the location of hot zones ten times within each cell line/tissue.

For local enrichment analysis, we optimised a method that was described in Schmitt et al. [55]. Briefly, for each interaction hot zone in each cell line/tissue, we obtained 250 bins upstream and 250 bins downstream, followed by calculating the enrichment score for each bin. The enrichment score is calculated by three steps, firstly, the expected enrichment bases is calculated by permuting the location of the hot zones ten times, then calculated the mean overlapping bases between the permuted hot zones and annotations for test. Then, the local enrichment score for each bin is calculated by the overlapping bases of observed bins divided by the expected overlapping bases, resulting in a total of 501 values, each of which are then divided by the minimum of these values to account for the magnitude of local enrichment. Finally, the local enrichment scores are log2 scaled followed by a z-score normalisation.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

See methods and supplementary documents.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

NL and JB developed and conceived the study in collaboration with HAR. NL wrote and developed approximately 90% of the manuscript, with editing provided by HAR and JB.

Acknowledgements

We acknowledged the publicly available Hi-C datasets and epigenome data.

References

1. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002;295:1306–11.

2. 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;326:289–93.

3. Rao SSP, 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;159:1665–80.

4. Ma W, Ay F, Lee C, Gulsoy G, Deng X, Cook S, et al. Using DNase Hi-C techniques to map global and local three-dimensional genome architecture at high resolution. Methods. 2018;142:59–73.

5. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. nature.com; 2016;17:333–51.

6. Ay F, Bailey TL, Noble W. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. 2014;24:999–1011.

7. 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:e0174744.

8. Cao Y, Chen Z, Chen X, Ai D, Chen G, McDermott J, et al. Accurate loop calling for 3D genomic data with cLoops. Bioinformatics. 2020;36:666–75.

9. Alinejad-Rokny H, Ghavami R, Rabiee HR, Rezaei N. MaxHiC: robust estimation of chromatin interaction frequency in Hi-C and capture Hi-C experiments. bioRxiv [Internet]. biorxiv.org; 2020; Available from: https://www.biorxiv.org/content/10.1101/2020.04.23.056226v1.abstract

10. Freedman ML, Monteiro ANA, Gayther SA, Coetzee GA, Risch A, Plass C, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011;43:513–8.

11. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.

12. Liu H, Luo X, Niu L, Xiao Y, Chen L, Liu J, et al. Distant eQTLs and Non-coding Sequences Play Critical Roles in Regulating Gene Expression and Quantitative Trait Variation in Maize. Mol Plant. 2017;10:414–26.

13. 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;47:598–606.

14. 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;167:1369–84.e19.

15. Greenwald WW, Chiou J, Yan J, Qiu Y, Dai N, Wang A, et al. Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. Nat Commun. 2019;10:2078.

16. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization [Internet]. arXiv [cs.LG]. 2014. Available from: http://arxiv.org/abs/1412.6980

17. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, et al. The European Nucleotide Archive. Nucleic Acids Res. 2011;39:D28–31.

18. Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, et al. The 4D nucleome project. Nature. 2017;549:219–26.

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;13:919–22.

20. Li G, Liu Y, Zhang Y, Kubo N, Yu M, Fang R, et al. Joint profiling of DNA methylation and chromatin architecture in single cells. Nat Methods. 2019;16:991–3.

21. Taberlay PC, Achinger-Kawecka J, Lun ATL, Buske FA, Sabir K, Gould CM, et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. Genome Res. 2016;26:719–31.

22. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27:849–64.

23. pairtools [Internet]. Github; [cited 2021 Mar 1]. Available from: https://github.com/open2c/pairtools

24. Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics. 2020;36:311–6.

25. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.

26. Si Z, Yu H, Ma Z. Learning Deep Features for DNA Methylation Data Analysis. IEEE Access. 2016;4:2732–7.

27. Chen X, Zhang B, Wang T, Bonni A, Zhao G. Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. BMC Bioinformatics. 2020;21:269.

28. Schölkopf B, Smola A, Müller K-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Comput. MIT Press; 1998;10:1299–319.

29. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker’s guide to Hi-C analysis: practical guidelines. Methods. 2015;72:65–75.

30. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.

31. SMIT, A. FA. Repeat-Masker Open-3.0. http://www.repeatmasker.org [Internet]. 2004 [cited 2021 Apr 20]; Available from: https://ci.nii.ac.jp/naid/10029514778/

32. Ward JH. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc. Taylor & Francis; 1963;58:236–44.

33. Branco MR, Pombo A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS Biol. 2006;4:e138.

34. Dekker J, Misteli T. Long-Range Chromatin Interactions. Cold Spring Harb Perspect Biol. 2015;7:a019356.

35. Maass PG, Barutcu AR, Rinn JL. Interchromosomal interactions: A genomic love story of kissing chromosomes. J Cell Biol. 2019;218:27–38.

36. Yasui D, Miyano M, Cai S, Varga-Weisz P, Kohwi-Shigematsu T. SATB1 targets chromatin remodelling to regulate genes over long distances. Nature. 2002;419:641–5.

37. Galande S, Purbey PK, Notani D, Kumar PP. The third dimension of gene regulation: organization of dynamic chromatin loopscape by SATB1. Curr Opin Genet Dev. 2007;17:408–14.

38. Bischof O, Purbey PK, Notani D, Urlaub H, Dejean A, Galande S, et al. Functional interaction between PML and SATB1 regulates chromatin-loop architecture and transcription of the MHC class I locus. Nat Cell Biol. Nature Publishing Group; 2007;9:45–56.

39. Frömberg A, Engeland K, Aigner A. The Special AT-rich Sequence Binding Protein 1 (SATB1) and its role in solid tumors. Cancer Lett. 2018;417:96–111.

40. Sunkara KP, Gupta G, Hansbro PM, Dua K, Bebawy M. Functional relevance of SATB1 in immune regulation and tumorigenesis. Biomed Pharmacother. 2018;104:87–93.

41. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

42. Carpten J, Nupponen N, Isaacs S, Sood R, Robbins C, Xu J, et al. Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. Nat Genet. 2002;30:181–4.

43. Casey G, Neville PJ, Plummer SJ, Xiang Y, Krumroy LM, Klein EA, et al. RNASEL Arg462Gln variant is implicated in up to 13% of prostate cancer cases. Nat Genet. 2002;32:581–3.

44. Anczuków O, Akerman M, Cléry A, Wu J, Shen C, Shirole NH, et al. SRSF1-Regulated Alternative Splicing in Breast Cancer. Mol Cell. 2015;60:105–17.

45. Klæstad E, Opdahl S, Engstrøm MJ, Ytterhus B, Wik E, Bofin AM, et al. MRPS23 amplification and gene expression in breast cancer; association with proliferation and the non-basal subtypes. Breast Cancer Res Treat. 2020;180:73–86.

46. Mahamdallie S, Yost S, Poyastro-Pearson E, Holt E, Zachariou A, Seal S, et al. Identification of new Wilms tumour predisposition genes: an exome sequencing study. Lancet Child Adolesc Health. 2019;3:322–31.

47. Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, et al. Expression Atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Res. academic.oup.com; 2018;46:D246–51.

48. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

49. Bunting KL, Melnick AM. New effector functions and regulatory mechanisms of BCL6 in normal and malignant lymphocytes. Curr Opin Immunol. 2013;25:339–46.

50. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710.

51. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010;107:21931–6.

52. Rada-Iglesias A. Is H3K4me1 at enhancers correlative or causative? Nat. Genet. 2018. p. 4–5.

53. Reddington JP, Perricone SM, Nestor CE, Reichmann J, Youngson NA, Suzuki M, et al. Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. Genome Biol. 2013;14:R25.

54. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. 2016;15:2038–49.

55. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep. 2016;17:2042–59.

56. Wit E de. TADs as the caller calls them. J Mol Biol [Internet]. 2019; Available from: http://dx.doi.org/10.1016/j.jmb.2019.09.026

57. Eres IE, Gilad Y. A TAD Skeptic: Is 3D Genome Topology Conserved? Trends Genet. 2021;37:216–23.

58. Oudelaar AM, Higgs DR. The relationship between genome structure and function. Nat Rev Genet. 2021;22:154–68.

59. Collins SJ, Gallo RC, Gallagher RE. Continuous growth and differentiation of human myeloid leukaemic cells in suspension culture. Nature. 1977;270:347–9.

60. Millius A, Weiner OD. Manipulation of neutrophil-like HL-60 cells for the study of

directed cell migration. Methods Mol Biol. 2010;591:147–58.

61. Yu H, Cook TJ, Sinko PJ. Evidence for diminished functional expression of intestinal transporters in Caco-2 cell monolayers at high passages. Pharm Res. 1997;14:757–62.

62. Hughes P, Marshall D, Reid Y, Parkes H, Gelber C. The costs of using unauthenticated, over-passaged cell lines: how much more data do we need? Biotechniques. 2007;43:575, 577–8, 581–2 passim.

63. Jin W, Penington CJ, McCue SW, Simpson MJ. A computational modelling framework to quantify the effects of passaging cell lines. PLoS One. 2017;12:e0181941.

64. Barshir R, Hekselman I, Shemesh N, Sharon M, Novack L, Yeger-Lotem E. Role of duplicate genes in determining the tissue-selectivity of hereditary diseases. PLoS Genet. 2018;14:e1007327.

65. Basha O, Argov CM, Artzy R, Zoabi Y, Hekselman I, Alfandari L, et al. Differential network analysis of multiple human tissue interactomes highlights tissue-selective processes and genetic disorder genes. Bioinformatics. 2020;36:2821–8.

66. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease [Internet]. Nature. 2011. p. 307–17. Available from: http://dx.doi.org/10.1038/nature10209

67. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Uncovering disease-disease relationships through the incomplete interactome. Science [Internet]. American Association for the Advancement of Science; 2015 [cited 2021 Apr 21];347. Available from: https://science.sciencemag.org/content/347/6224/1257601.abstract?casa_token=u10-zC1F6-QAAAAA:8I4Xw7kaOr77iyYTRwp84_8KdA-uR1g6dzScCxh8ROITcIe4CI8NOaJrg7uYsmpqLTx_vT2OuLZo2-k

68. Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nat Methods. 2016;13:366–70.

69. Hekselman I, Yeger-Lotem E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. Nat Rev Genet. 2020;21:137–50.

70. Ibrahim DM, Mundlos S. The role of 3D chromatin domains in gene regulation: a multi-facetted view on genome organization. Curr Opin Genet Dev. 2020;61:1–8.

71. 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;485:376–80.

72. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161:1012–25.

73. Smith EM, Lajoie BR, Jain G, Dekker J. Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the

CFTR Locus. Am J Hum Genet. 2016;98:185–201.

74. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science. 2016;351:1454–8.

75. Aspera - Overview [Internet]. [cited 2021 Apr 21]. Available from: https://www.ibm.com/products/aspera

76. Andrews S, Others. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

77. Ward CM, To T-H, Pederson SM. ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files. Bioinformatics. 2020;36:2587–8.

78. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016;9:88.

79. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv [q-bio.GN]. 2013. Available from: http://arxiv.org/abs/1303.3997

80. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. JMLR. org; 2011;12:2825–30.

81. 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;17:261–72.

82. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16:259.

83. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res. 2016;44:e70.

84. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. 2019;47:D853–8.

Supplementary information

Note: Supplementary Tables and Files are hosted on Figshare:

Supplementary Table 1-4, 6-10: https://figshare.com/s/aa608477a3248e062fe0

Supplementary Table 5: https://figshare.com/s/ff27719c21078d85b745

Supplementary File 1: https://figshare.com/s/72fdd11eaee80758c719

Supplementary File 2: https://figshare.com/s/1899f248ffbbd5110b6a

Supplementary Table 1: Metadata of all curated Hi-C datasets in this study.

Supplementary Table 2: The information of 196 Hi-C datasets after filtering of sequencing read count.

Supplementary Table 3: Statistics of the analysis of Hi-C datasets.

Supplementary Table 4: The information of 51 cell lines and tissues and their Hi-C datasets.

Supplementary Table 5: Repeat regions of hg38 genome used in this study.

Supplementary Table 6: MADSSI count of 51 cell lines and tissues.

Supplementary Table 7: Top10 common interactions found across 51 cell lines and tissues.

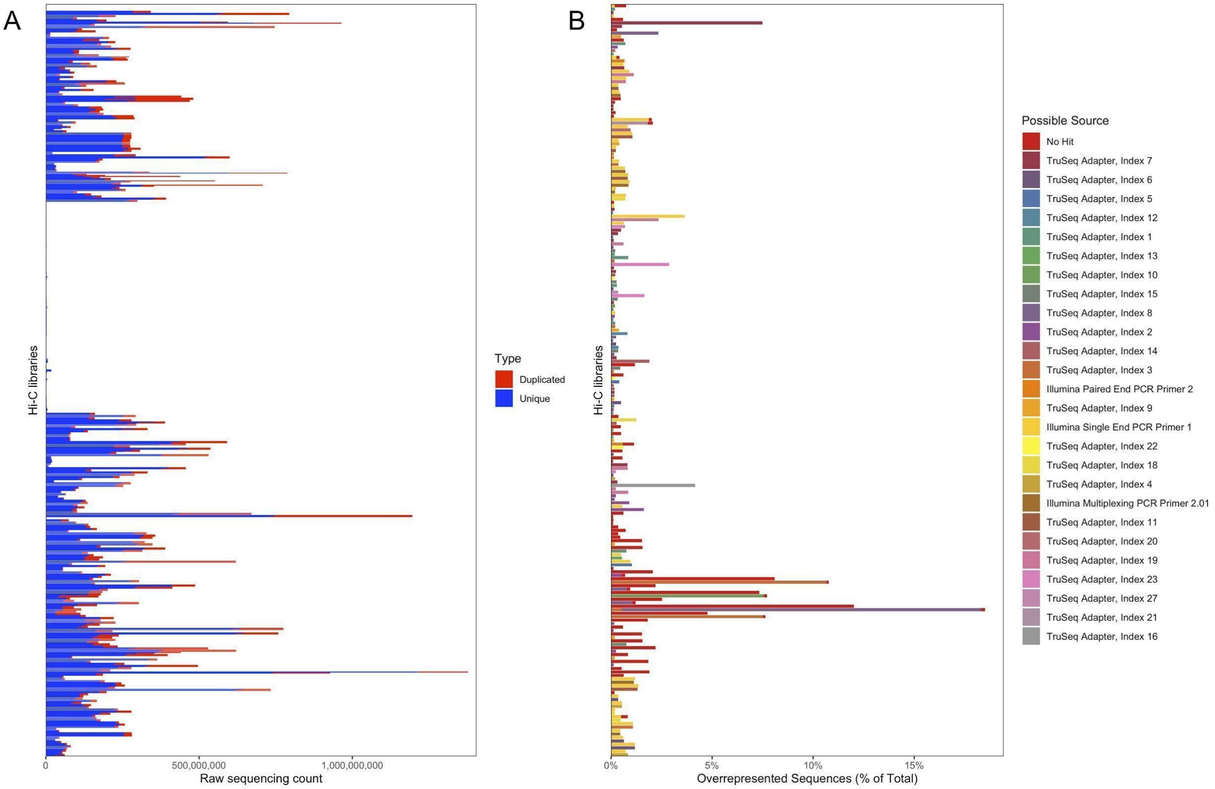
Supplementary Table 8: Normalised (log2-scaled) gene expression profile of genes contacted by top 10 non-unique interactions in different cells and tissues from EBI-Expression Atlas.

Supplementary Table 9: Gene ontology over-representation analysis of genes contacted by cell lines/tissues-unique MADSSI.

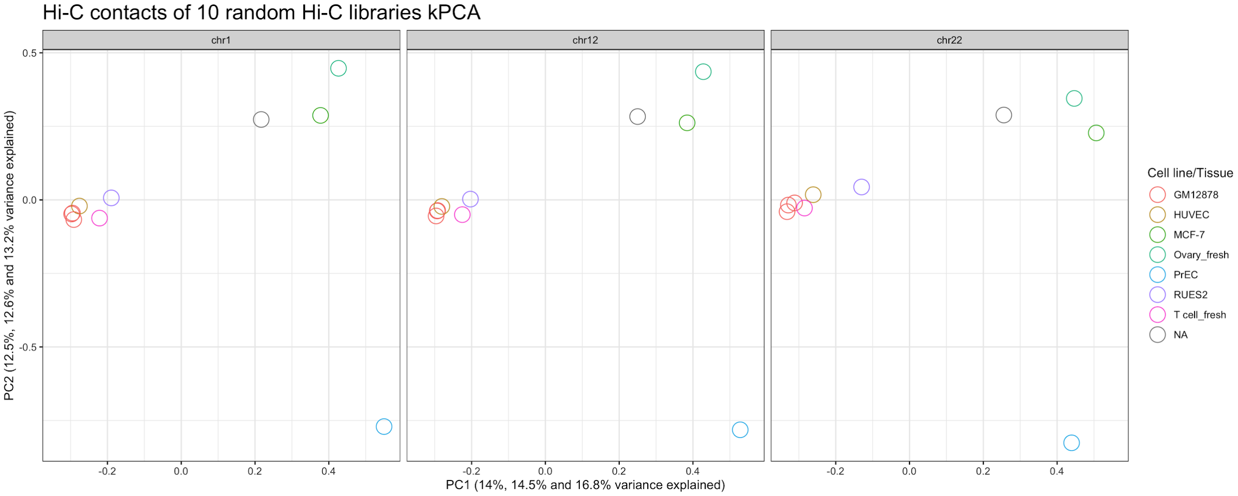
Supplementary Table 10: The identified interaction hotzones across 51 cell lines and tissues.

Supplementary File 1: MaxHiC-detected statistically significant interactions (MADSSI) identified for 51 cell lines and tissues.

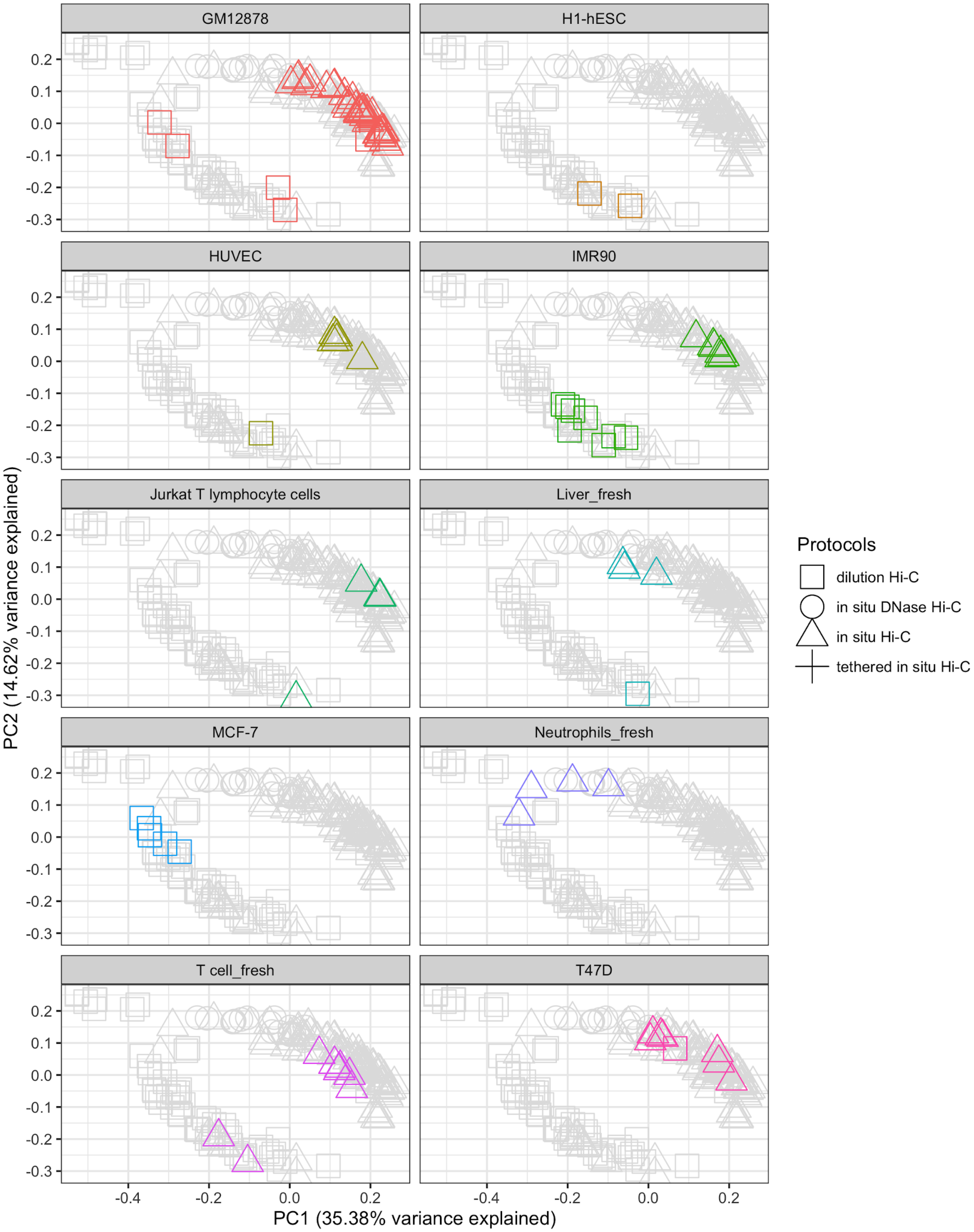
Supplementary File 2: Topologically-associated domains (TADs) identified in each cell line/tissue.



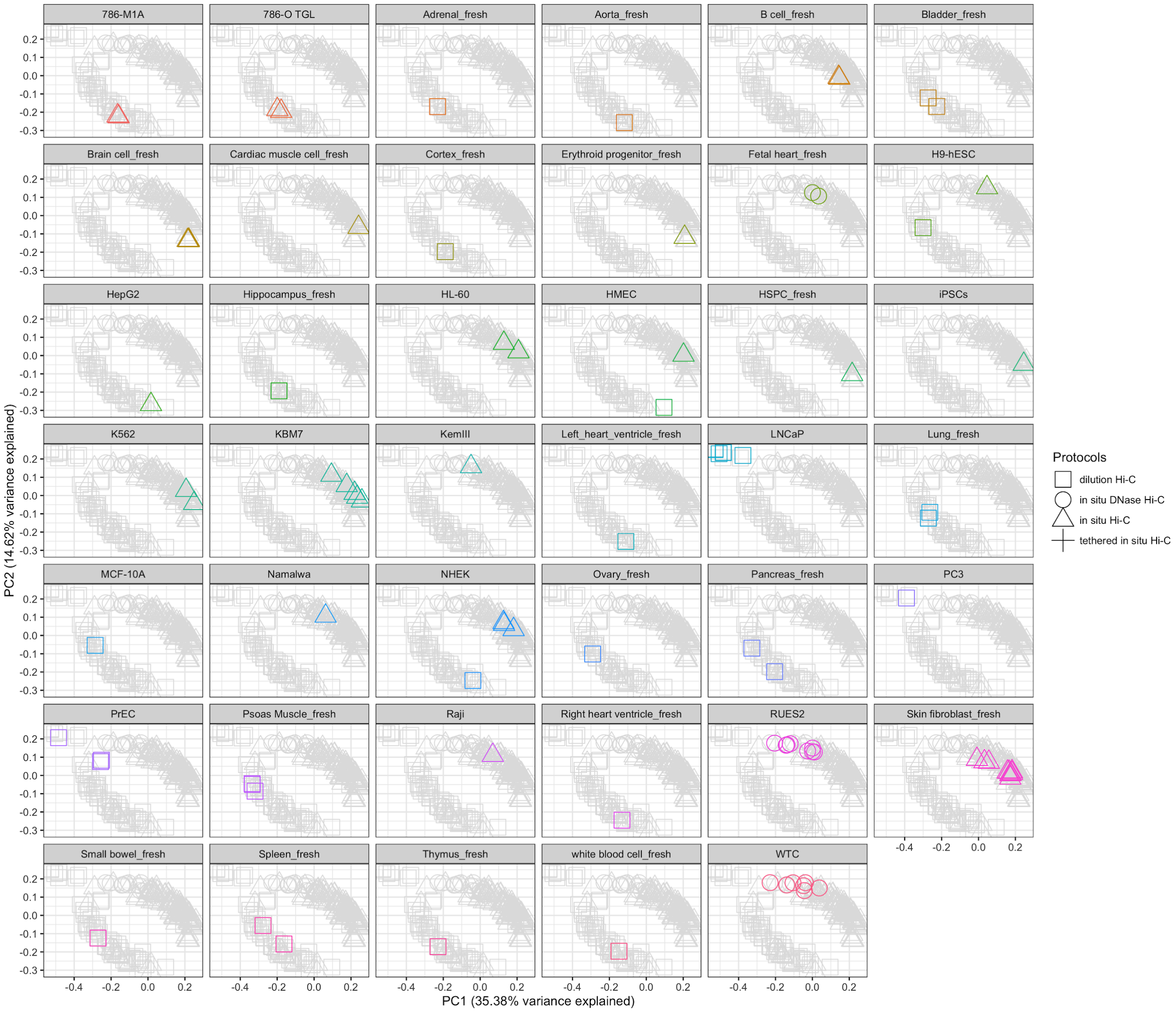
Supplementary Figure 1: Summary of FastQC reports generated using *ngsReport* [77]. A: Total read count of each Hi-C library. B: Overrepresented sequences of each Hi-C library.



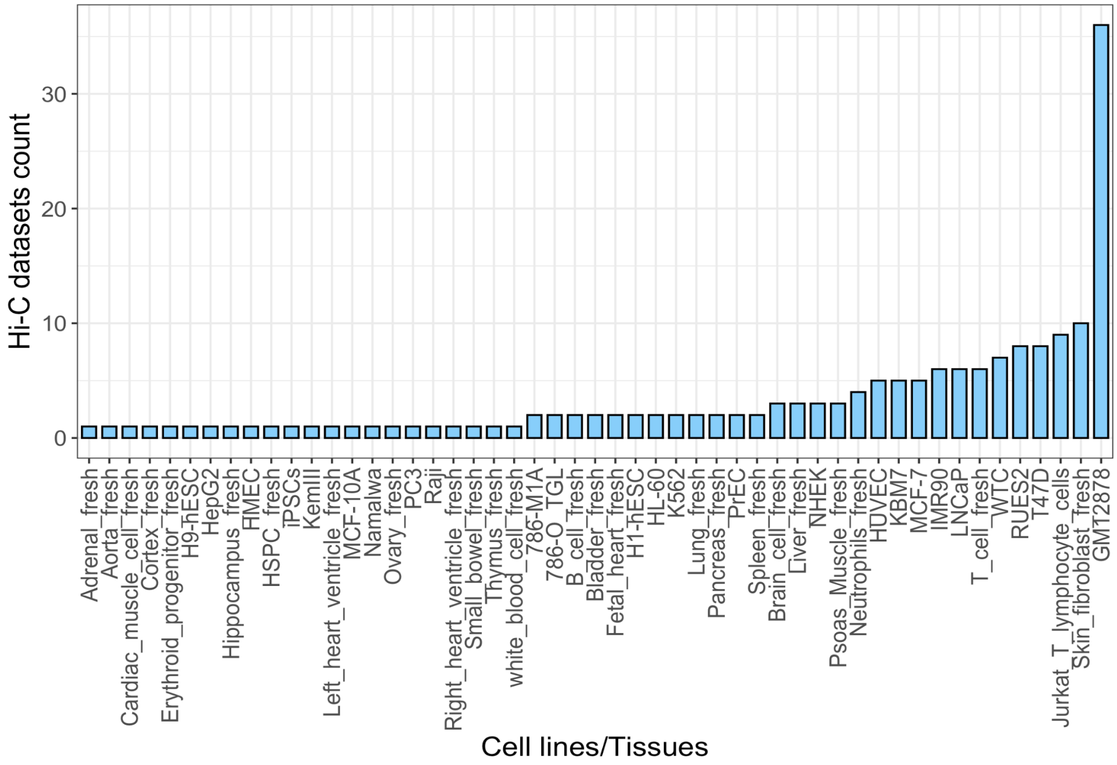
Supplementary Figure 2: Kernel principal component analysis of intra-chromosomal interactions of 10 random Hi-C libraries on chromosome 1, 12 and 22.



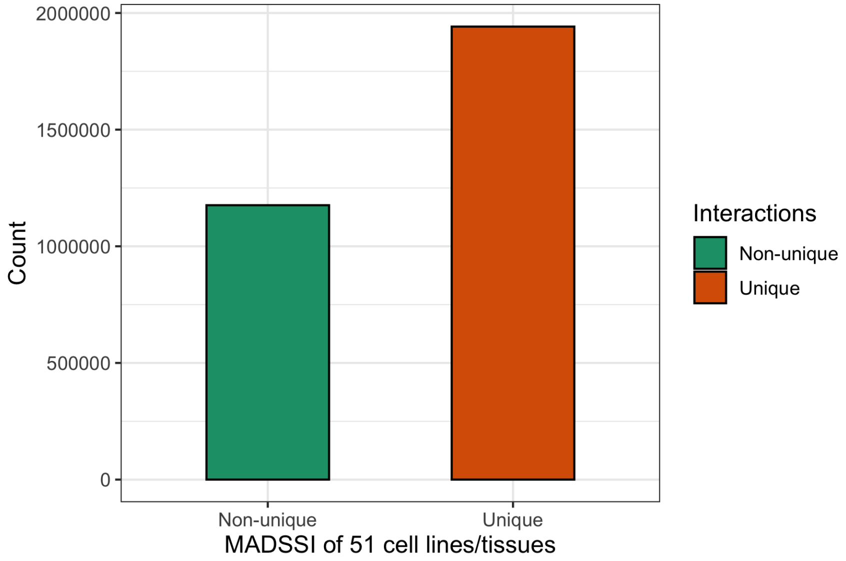
Supplementary Figure 3: Kernel principal component analysis of intra-chromosomal interactions of 196 Hi-C libraries. Hi-C libraries that belong to cell lines or tissues that were sampled in more than one study are colored, while other libraries are colored with grey.



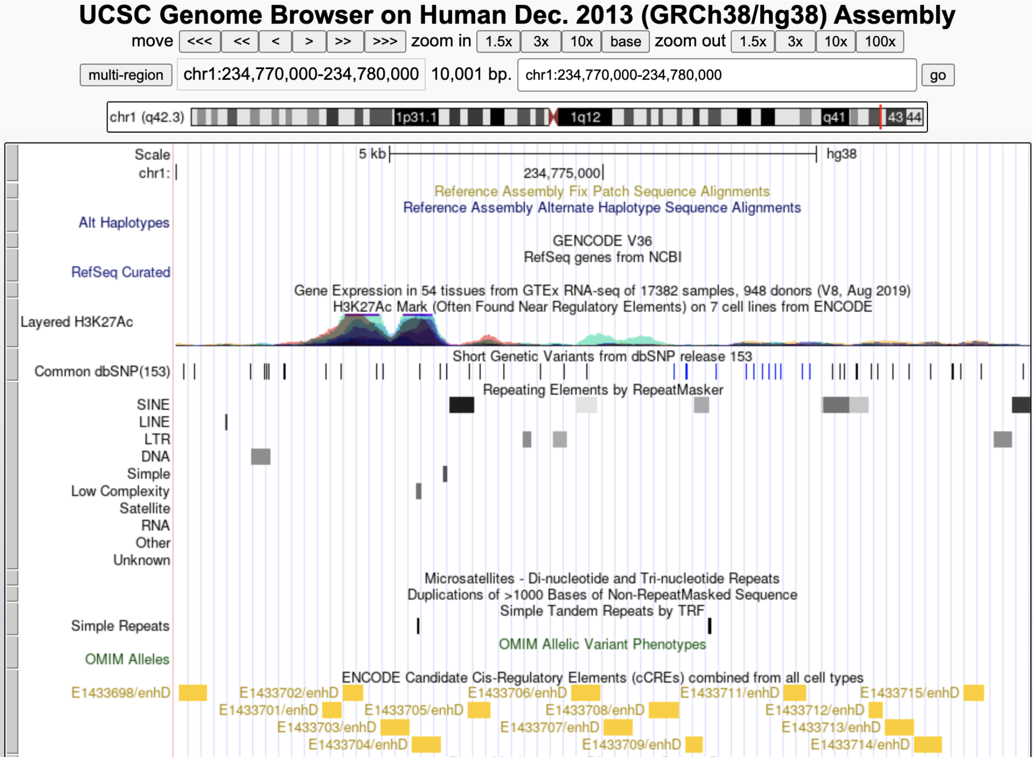
Supplementary Figure 4: Kernel principal component analysis of intra-chromosomal interactions of 196 Hi-C libraries. Hi-C libraries that belong to cell lines or tissues that were sampled in only one study are colored, while other libraries are colored with grey.



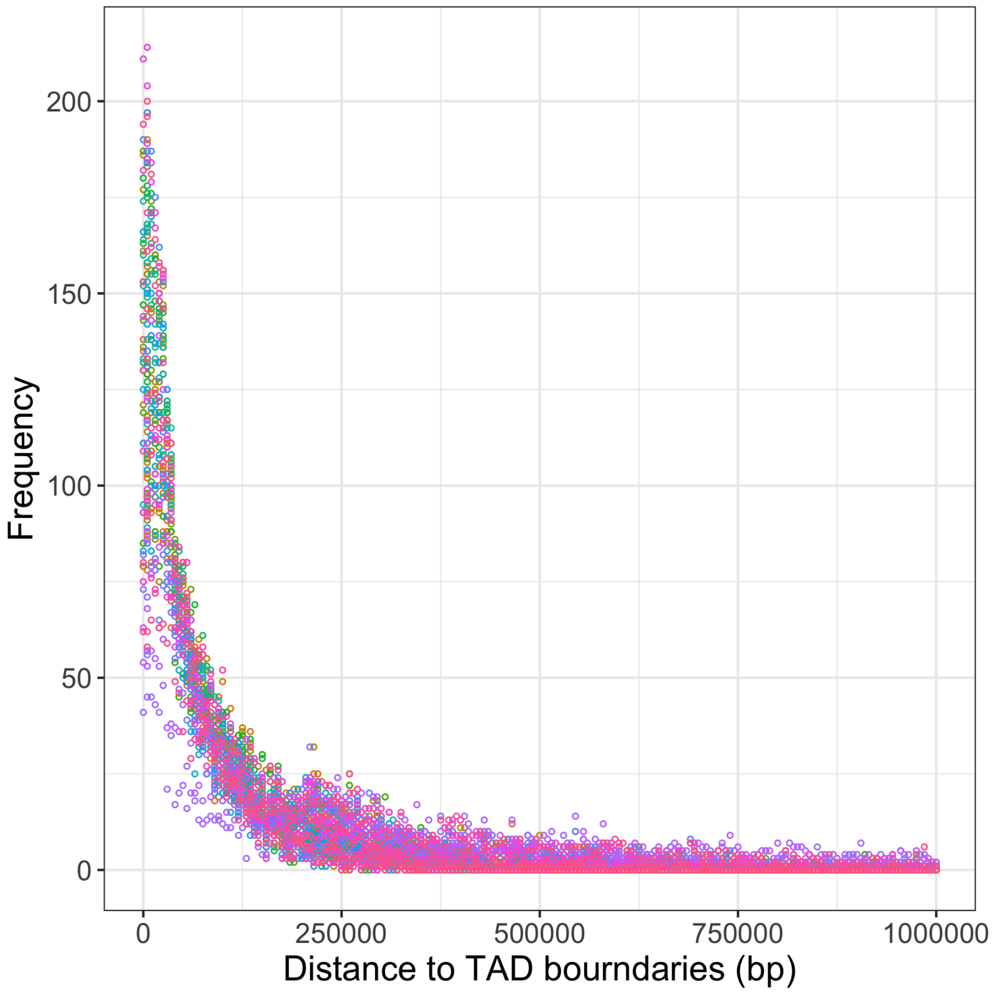
Supplementary Figure 5: Hi-C libraries count per cell line/tissue that used to generate cell line/tissue-specific informative read pairs.



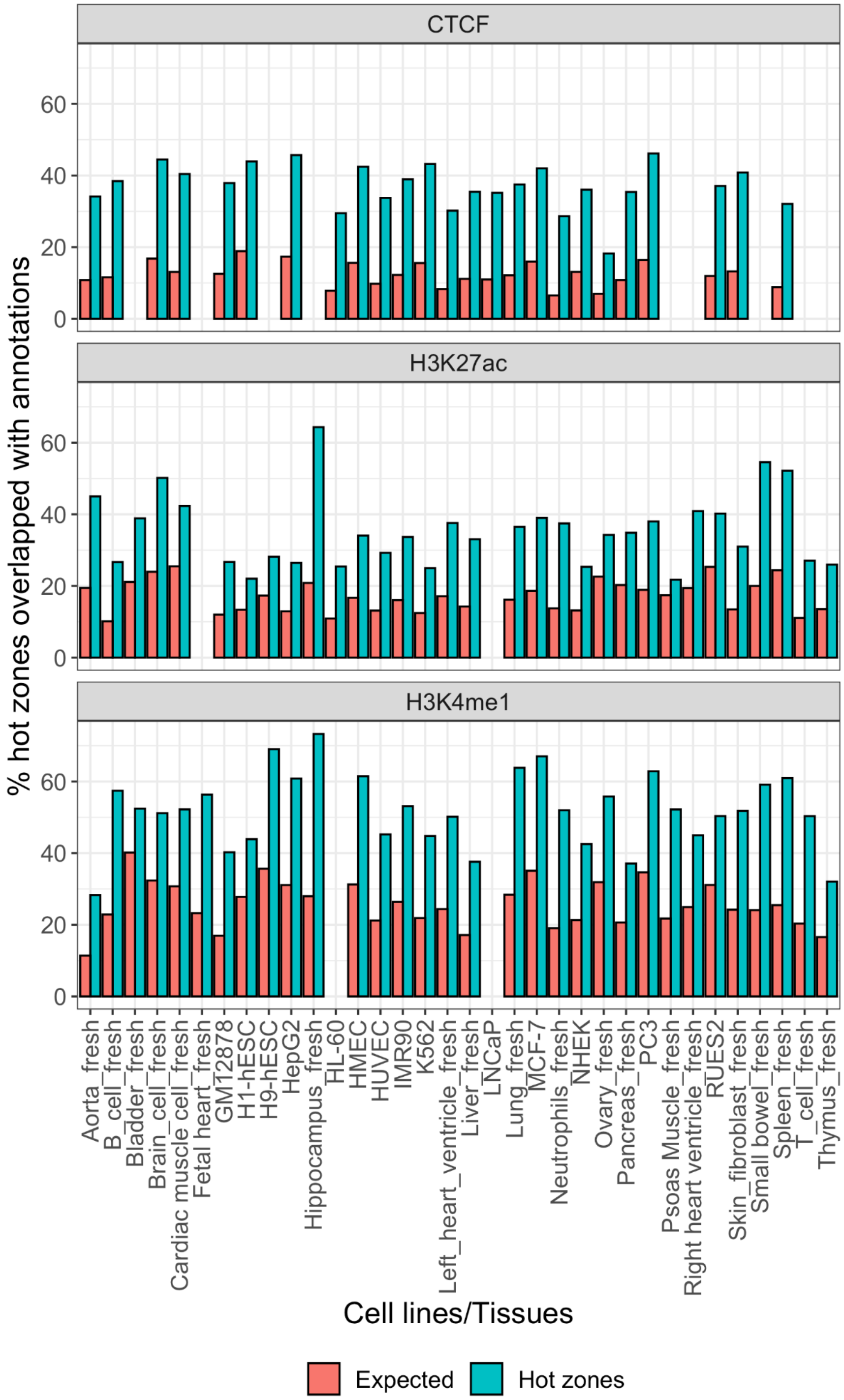
Supplementary Figure 6: Count of unique and non-unique MaxHiC-detected statistically significant interactions (MADSSI) across 51 cell lines/tissues.


UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region chr1:234,770,000-234,780,000 10,001 bp. chr1:234,770,000-234,780,000 go
chr1 (q42.3) 1p31.1 1q12 q41 43 44
Scale
chr1:
5 kb
hg38
234,775,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
Alt Haplotypes
GENCODE V36
RefSeq genes from NCBI
RefSeq Curated
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
Short Genetic Variants from dbSNP release 153
Common dbSNP(153)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
Microsatellites - Di-nucleotide and Tri-nucleotide Repeats
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Simple Tandem Repeats by TRF
Simple Repeats
OMIM Allelic Variant Phenotypes
OMIM Alleles
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
E1433698/enhD
E1433702/enhD
E1433706/enhD
E1433711/enhD
E1433715/enhD
E1433701/enhD
E1433705/enhD
E1433708/enhD
E1433712/enhD
E1433703/enhD
E1433707/enhD
E1433713/enhD
E1433704/enhD
E1433709/enhD
E1433714/enhD

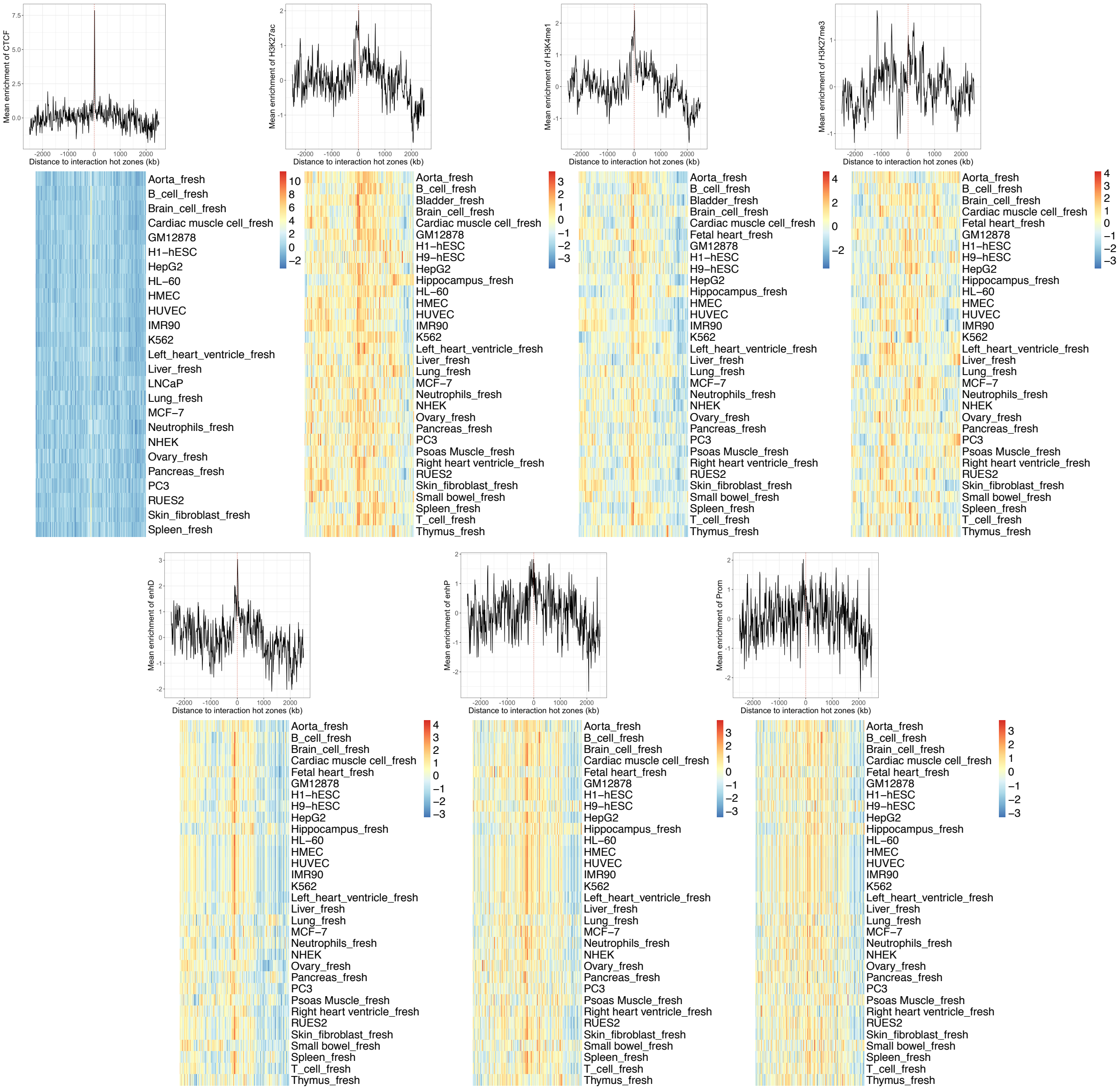
Supplementary Figure 7: A screenshot of the region chr1:234,770,000-234,780,000 in the UCSC genome browser [84].



Supplementary Figure 8: Distribution of relative distance between interaction hot zones and TAD boundaries. Colors indicate TAD boundaries in different cell lines/tissues, color legend is hidden for better visualisation.



Supplementary Figure 9: Global enrichment analysis with interaction hot zones. Bar plot shows the overlap (green) between interaction hot zones and CTCF-binding sites, H3K27ac and H3K4me1 histone markers annotated in different cell lines/tissues. The expected overlaps (red) are calculated by permuting the location of interaction hot zones of each cell line/tissue and calculating the overlaps. The y-axis is the percentage of interaction hot zones overlapped with corresponding annotations.



Supplementary Figure 10: Local enrichment analysis with interaction hot zones. Local enrichment of CTCF-binding sites, H3K27ac, H3K4me1, H3K27me3, cCRE distal enhancer signatures, cCRE proximal enhancer signatures and cCRE promoter signatures, centering on interaction hot zones in each cell line/tissue. Top panel: mean of enrichment score of a 5 Mb region centered at interaction hot zones across cell lines/tissues. Bottom panel: enrichment score of a 5 Mb region centered at interaction hot zones in each cell line/tissue.

Chapter 5

Identifying human cell/tissue type-specific potentially functional interactions

Identifying human cell/tissue type-specific potentially functional interactions

Ning Liu[1,2,3], Hamid Alinejad-Rokny[4*], James Breen[1,2,3,5*]

[1] South Australian Health & Medical Research Institute, Adelaide, Australia
[2] Robinson Research Institute, University of Adelaide, Adelaide, Australia
[3] Adelaide Medical School, University of Adelaide, Adelaide, Australia
[4] Systems Biology and Health Data Analytics Lab, The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, NSW, 2052, Australia
[5] South Australian Genomics Centre (SAGC), Adelaide, Australia

* corresponding authors

Contributions: NL and JB developed and conceived the study in collaboration with HAR. NL wrote and developed approximately 90% of the manuscript, with editing provided by HAR and JB. Manuscript is formatted for submission to *Genome Biology*.

Abstract

Background

The three-dimensional chromosome structure allows physical engagement between distantly located DNA fragments, which can affect gene regulation. High resolution chromosome conformation capture (Hi-C) sequencing data offers the opportunity to identify interactions across the genome. Previously, we cataloged MaxHiC-detected statistically significant interaction (MADSSI) profiles for 51 human cell and tissue types using 173 published Hi-C datasets and a statistical background model implemented in MaxHiC. However, the potential biological function of the chromatin interactions identified within this set have yet to be investigated.

Results

In this study, we cataloged 66 different interaction classes across cell line/tissue-specific MADSSI via epigenomics data integration contained within public databases such as FANTOM5, Epigenomics Roadmap and ENCODE. We successfully annotated an average of 75.35% of MADSSI across all cells and tissues, generating a comprehensive annotation of cell/tissue-specific interaction profiles. Focusing on interactions that are more likely to be regulatory functional, we generated lists of potentially regulatory functional MADSSI (PROF-MADSSI). While interactions can be annotated to multiple interaction types, PROF-MADSSI were dominated by enhancer-enhancer interactions (69.6% on average), enhancer-CTCF interactions (49.6% on average) and promoter-enhancer interactions (21.5% on average). Finally, we profiled cell/tissue-specific 3D regulatory regions, defined as specific regions with regulatory elements contacting cell/tissue-specific expressed gene promoters. We found 3D regulatory regions are enriched for the type-specific super-enhancers, and on average 26.73% of them overlapped with tissue-specific eQTLs, demonstrating their importance in the cell/tissue type-specific gene expression regulation.

Conclusion

We present a comprehensive database of annotated cell/tissue-specific MADSSI profiles, identified potentially functional chromatin interactions and regions, through 3D structure. This information can facilitate future genetics research of complex gene regulations in specific cell and tissue types, providing essential context to non-coding DNA regions in human genetics studies.

Keywords

Chromatin interactions; Hi-C; 3D chromosome structure.

Background

The Hi-C sequencing assay has been widely used to investigate the three-dimensional (3D) chromosome architecture and detect chromatin interactions between functional components of the cellular genetic system [1–3]. Chromosome interactions at large and small resolution have been identified and defined using Hi-C and related chromosome capture sequencing approaches, starting from macro-level structures such as A/B compartments [4] and topologically-associated domains (TADs) [5], to more high resolution micro-level structures such as chromatin loops [6] and frequently interacting regions (FIREs) [7].

The investigations of these chromosome features have revealed that the 3D chromatin interactions play important roles in the regulation of genes, bringing distal regulatory elements such as enhancers into close proximity with gene promoters to facilitate gene initiation [8], or forming polycomb-bound complex loops to mediate gene repression [9]. Despite its essential role in cellular function, non-specific physical interactions found within the nucleus make it difficult to accurately detect biologically functional interactions from hundreds/millions/billions of uninformative Hi-C reads. Additional technical issues, such as amplification biases across the genome during sequencing, variation in genomic fragment size of each interaction due to various density of enzyme cutting sites, and the occurrence of random interactions caused by random looping or ligation artefacts [10,11] also impact the ability to identify relevant physical interactions.

To address this challenge, a number of methods have been developed to detect interactions that are more likely to be authentic and biologically relevant, using statistical models to identify statistically significant interactions from raw Hi-C contacts [10–13]. In a previous study, our collaborators developed MaxHiC, which outperforms other methods in identifying statistically significant interactions involving regulatory features [14]. Using MaxHiC to selectively reduce the relevant interactions for analysis, we established an analysis pipeline based on the 4DN Hi-C analysis pipeline [15] and generated MaxHiC-detected statistically significant interaction (MADSSI) profiles across 51 human cell lines and tissues collected from publicly available Hi-C data repositories such as European Nucleotide Archive (ENA) [16] and the 4DN data portal [15].

In order to annotate Hi-C chromatin interactions to reveal their potential biological functions, epigenomic and tissue-specific annotations (i.e. histone modification markers, promoters, enhancers and repressors) have been used to provide context to each interaction [2,17–20]. In this study, we annotated cell/tissue-specific MADSSI profiles with cell/tissue-specific epigenomics annotations and classified them into 66 classes of chromatin interactions. Subsequently, we investigated the interaction classes which are potentially regulatory functional. Finally, looking at interaction classes involved in contacting cell/tissue-specific expressed gene promoters, we identify cell/tissue-specific 3D regulatory regions and integrate them with tissue-specific expression quantitative trait loci (eQTLs).

Results

Categories of statistically significant chromatin interactions

We analysed 173 published Hi-C datasets from the European Nucleotide Archive (ENA) [16] and the 4DN data portal [15], generating MADSSI profiles for 51 cell lines and tissues using a customised pipeline in our previous study. In order to comprehensively annotate MADSSIs identified in each cell line/tissue, we first categorised chromatin interactions into 66 interaction classes by overlapping chosen genomics annotations and each anchor bin in any MADSSI (Figure 1). The genomic annotations are obtained from four different sources. Firstly, cell/tissue type-specific expression data was obtained from the EBI-Expression Atlas [21]. Expression data was then integrated with known gene annotation from the GENCODE gene annotation database [22] to identify expressed genes and non-expressed genes, with promoters being defined by 2 kb regions upstream of the transcription start sites of genes (Figure 1). In order to define active enhancers, we used the cell/tissue type-specific expressed enhancers data defined by the Cap analysis gene expression (CAGE) sequencing data generated in the FANTOM5 project [23]. We also included human cell/tissue type-specific chromHMM states, which were systematically predicted by the Roadmap Epigenomics Project [24] using data to develop a 15-states chromHMM model [25] defining regulatory elements within each tissue. The cell/tissue specific chromHMM states data is used to define poised promoters, non-expressed active enhancers, bivalent enhancers, repressed polycomb, heterochromatin and repeat regions in each cell/tissue (Figure 1). Finally, CTCF-binding sites were obtained from the ENCODE project [26] and the CTCFBSDB [27]. Overall, using each annotation as an anchor point and looking at combinations of genomic interactions, we analysed 66 classes of chromatin interactions throughout the study. For the consistency of annotation comparisons, we excluded cells or tissues that had too few annotations resulting in a total of 35 human cell and tissue types (Supplementary Table 1).

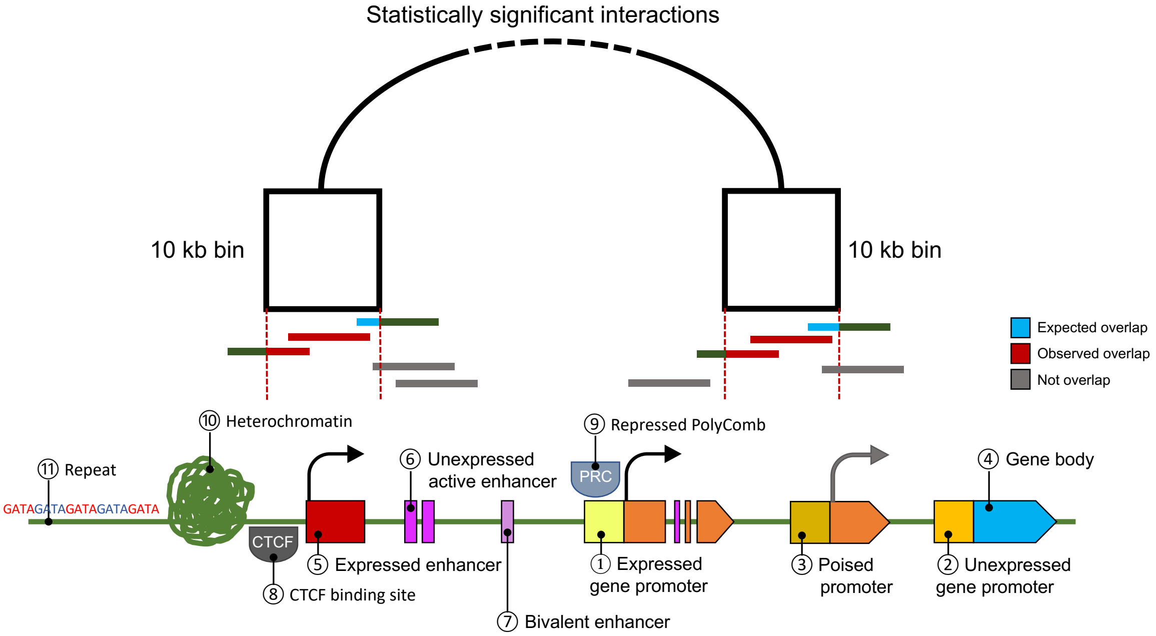
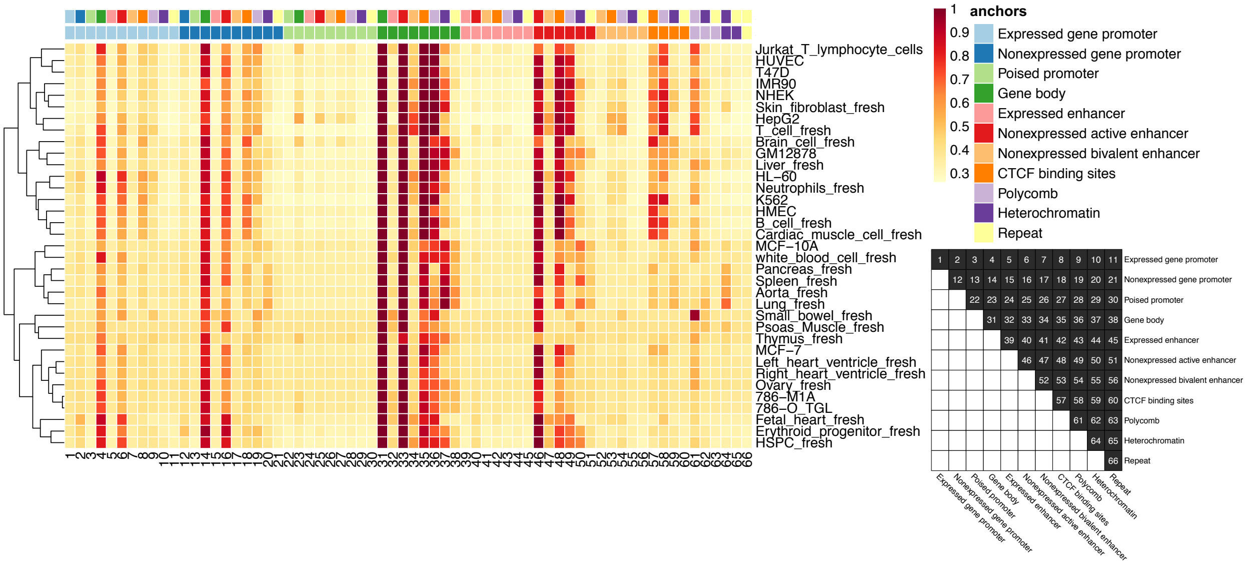


Figure 1: Schematics overview of definition of interaction classes. Interaction classes are defined by the overlapping between chosen annotations and both anchor bins of statistically significant interactions (top panel). 11 types of chosen annotations are described in the schematic figure (bottom panel).



1	2	3	4	5	6	7	8	9	10	11	Expressed gene promoter
	12	13	14	15	16	17	18	19	20	21	Nonexpressed gene promoter
		22	23	24	25	26	27	28	29	30	Poised promoter
			31	32	33	34	35	36	37	38	Gene body
				39	40	41	42	43	44	45	Expressed enhancer
					46	47	48	49	50	51	Nonexpressed active enhancer
						52	53	54	55	56	Nonexpressed bivalent enhancer
							57	58	59	60	CTCF binding sites
								61	62	63	Polycomb
									64	65	Heterochromatin
										66	Repeat
Expressed gene promoter	Nonexpressed gene promoter	Poised promoter	Gene body	Expressed enhancer	Nonexpressed active enhancer	Nonexpressed bivalent enhancer	CTCF binding sites	Polycomb	Heterochromatin	Repeat	

Figure 2: The normalised count of each interaction class annotated in each cell and tissue. Heatmap (left) shows the number of interaction classes identified in each cell/tissue. Each column corresponds to an interaction type id. Anchors indicated the annotations in each anchor of an interaction to define an interaction class. Value in each cell is the percentile rank of the z-score normalised count of the interaction type in each row (cell/tissue). Heatmap (right) shows the 66 interaction classes.

On average, approximately 75.35% of MADSSI in each cell/tissue were annotated by both anchor bins overlapping with selected genomic elements (Supplementary Figure 1 and Supplementary File 1). To compare the frequency of interaction classes across all cells/tissues while minimising the bias of various MADSSI count in different cells/tissues (Supplementary Figure 2), we normalised the frequency of interaction classes found in each cell/tissue using z-score normalisation and calculated the rank percentile to demonstrate the frequency of interaction classes found in each cell/tissue. We found out that gene bodies (Class 4, 14, 31, 33, 35 and 36) and non-expressed active enhancers (Class 6, 16, 46, 48 and 49) interaction classes were two of the most frequently classes found in the majority of cells/tissues (Figure 2). We also found poised promoter (Classes 21 - 30) and expressed enhancer (Classes 39 - 45) interaction classes have relatively low frequency across all 35 cells and tissues. This may be due to these two annotations having relatively smaller effect size, which is calculated by the production of the average length and total count of the annotation (Supplementary Figure 3). In addition to enhancer-related interaction classes, which are associated with gene activation, we also observed frequent interactions of CTCF binding sites and overlapping polycomb anchors (Classes 57, 58 and 61) in some cells and tissues, including B cells, brain cells, cardiac muscle cells, HepG2, K562, NHEK, skin fibroblast and T cells (Figure 2), indicating that some MADSSI in these cells and tissues are also associated with gene repression via polycomb-mediation. By clustering of the similar frequency pattern of the annotated interaction classes, we found cell/tissue types with biological similarity are clustered together, such as HSPC and erythroid progenitor cells, 786-M1A and 786-O_TGL cell lines, left heart ventricle and right heart ventricle tissues, neutrophils and HL-60 cell lines (Figure 2), indicating that similar cell types tend to have similar annotated MADSSI profiles.

Potentially regulatory functional chromatin interactions

In the 66 interaction classes analysed above, some interaction classes are important to demonstrate extra layers of regulatory mechanisms of gene expression, such as expressed gene promoter contacting enhancers, CTCF binding sites contacting enhancers and polycomb contacting expressed gene promoters. We then drilled down to interactions that are potentially regulatory functional given our hypothesis of expressed gene promoters and enhancers being potential regulatory markers to indicate cell/tissue-specific functional regulations, defining 6 types of potentially regulatory functional MADSSI (PROF-MADSSI) for each cell/tissue type (Table 1 and Figure 3A). We describe promoter-promoter interactions (PPI) as contacts between two promoters of cell/tissue-specific expressed genes or contacts between unexpressed gene promoters and expressed gene promoters (Figure 3A). The former class can lead to the formation of a cooperatively transcribed network among genes [18], while the latter may be regulatory when gene promoters act as enhancers to regulate expression of other genes [28–30].

Table 1: Definition of potentially regulatory functional MADSSI (PROF-MADSSI). Annotations 1 and 2 indicate the annotation used in each anchor of an interaction to define a PROF-MADSSI.

Annotation 1	Annotation 2	Potential regulatory function	Type
Expressed gene promoter	Expressed gene promoter	Co-expression; promoter acting as an enhancer	PPI
	Non-expressed gene promoter	Promoter acting as an enhancer	
	Expressed enhancer	Enhancer enhancing gene expression	PEI
	Non-expressed active enhancer		
	Bivalent enhancer		
	CTCF binding sites	CTCF drive PEI or PPI	PCI
	Polycomb	Polycomb mediate gene expression	PPCI
Expressed enhancer	Expressed enhancer	Enhancer-enhancer interaction	EEI
	Non-expressed active enhancer		
	Bivalent enhancer		
	CTCF binding sites	CTCF promote promoter EP or PP contacts	ECI
Non-expressed active enhancer	Non-expressed active enhancer	Enhancer-enhancer interaction	EEI
	Bivalent enhancer		
	CTCF binding sites	CTCF promote promoter EP or PP contacts	ECI
Bivalent enhancer	Bivalent enhancer	Enhancer-enhancer interaction	EEI
	CTCF binding sites	CTCF drive PEI or PPI	ECI
Expressed gene promoter	Poised promoter	Co-expression; promoter acting as an enhancer	PPI

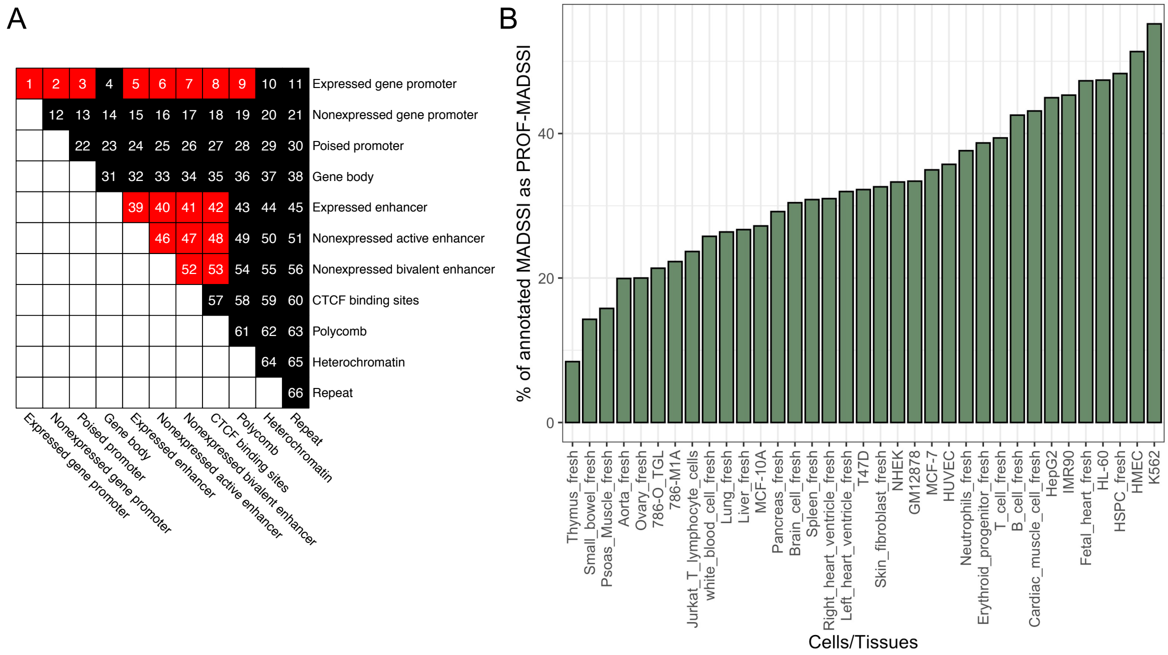


Figure 3: The identification of PROF-MADDSI. (A) Heatmap shows all interaction classes based on the anchored elements, interaction classes colored in red are selected to be PROF-MADSSI. (B) Bar plot shows the percentage of annotated MADSSI identified as PORF-MADSSI in 35 cells/tissues, x-axis is ordered as supplementary figure 2.

As a classical regulation model [2,8,31], promoter-enhancer interactions (PEI) which could be interactions between expressed gene promoters or poised promoters in one side and expressed enhancers, non-expressed active enhancers and bivalent enhancers in another side, are shown to be regulatory functional [2,8,32]. Similarly, enhancer-enhancer interactions (EEI), which can lead to cooperative regulation between enhancers [32–34] are considered as PROF-MADSSI as well. In addition to the well-established mechanism where the chromosome insulator protein CTCF forms the basis of chromatin loops [35,36] and the boundaries of topologically-associated domains [5,6], the binding of CTCF in promoter regions can also drive promoter-promoter and promoter-enhancer interactions [37]. Therefore, promoter-CTCF binding site interactions

(PCI) and enhancer-CTCF binding site interactions (ECI) are also regarded as PROF-MADSSI. Finally, the polycomb complex can mediate transcription by changing the accessibility of DNA, by affecting the function of RNA polymerase when binding to gene promoters and by involvement in promoter-promoter interactions [38–40], all reasons why the promoter-polycomb interactions (PPCI) are included for their potential regulatory function.

Overall, we identified on average 32.81% of MADSSI as PROF-MADSSI in 35 cells/tissues (Supplementary File 2), with thymus being the tissue that have the least fraction (8.4%) of PROF-MADSSI from its annotated MADSSI and K562 being the most (55.2%) (Figure 3B). We observed that the vast majority of all PROF-MADSSI are composed of enhancers-related interactions, including on average 69.6% of them are EEI, 49.6% of them are ECI and 21.5% of them are PEI (Figure 4A). Consistently, we found that EEI and ECI are the two most frequently detected interaction types across 35 cells/tissues (Figure 4B). We observed that 35 cells/tissues are clustered into two large clusters, one have more ECI than EEI, such as Brain, NHEK and B cells, the other one have more EEI than ECI, such as thymus, small bowel and psoas muscle (Figure 4B). Interestingly, we also found that all tissues except for liver (11 of 12), have more EEI than ECI (Figure 4B).

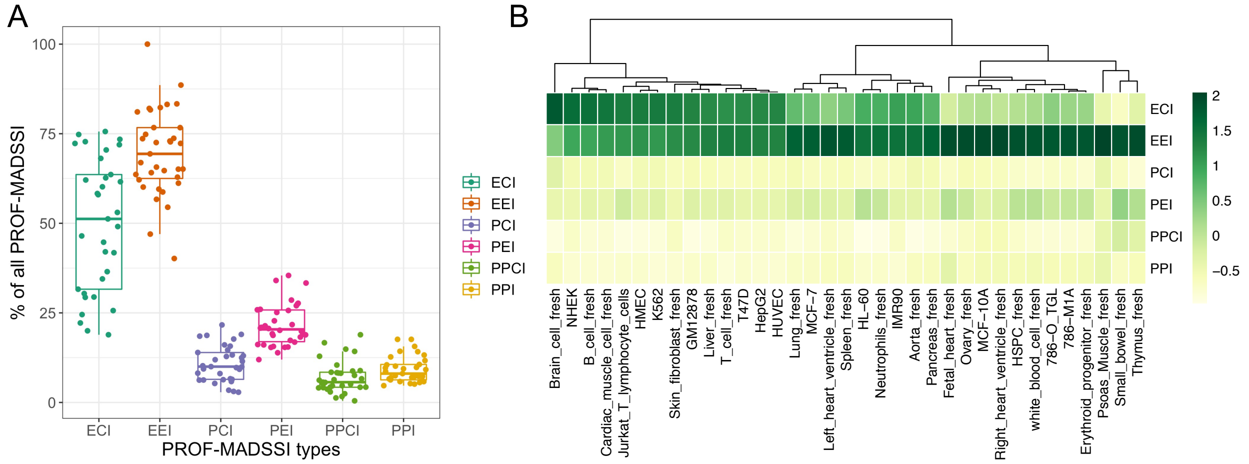


Figure 4: Composition of the identified PROF-MADSSI. (A) Box plot shows the percentage of six PROF-MADSSI types in all cells and tissues. (B) Heatmap shows the frequency of each PROF-MADSSI types across 35 cells/tissues. Each cell in the heatmap is a z-score normalised count.

We hypothesised that if chromatin interactions are assigned to more than two PROF-MADSSI types (i.e. EEI, PPI, PEI etc), such interactions are stacked with regulatory elements, potentially implying biological functionality. We therefore looked into the PROF-MADSSI of each cell/tissue, we found on average 42% of the interactions are assigned to at least two PROF-MADSSI types in all cells/tissues (Figure 5A). Taking two extreme cases as examples, 6 PROF-MADSSI were detected in Psoas Muscle Tissue and all of them are annotated as EEI, while of the 183,346 PROF-MADDSI detected in K562 cells, 61.9% of them are annotated with at least two PROF-MADSSI types (Figure 5B). Interestingly, in the Leukemia cell line K562, 1,416 chromatin interactions are annotated as all 6 PROF-MADSSI types (i.e. EEI, PEI, PPI, ECI, PCI and PPCI) (Figure 5B), indicating regulatory annotations including enhancers, expressed gene promoters, CTCF binding sites and polycomb signals are captured by these interactions. For chromosome 12 in K562, a PROF-MADSSI contacts with the promoter region (2 kb upstream of transcription start site) of gene *STAT6*, which encodes transcription factor STAT6 associated with the development of lymphoma and leukemia [41]. We also observed CTCF-binding sites overlaid in the interacting regions, along with the predicted active enhancer states (red) and polycomb-complex states (gray) based on K562 chromHMM states (Figure 6). This suggests that this identified PROF-MADSSI may point to regulation of the *STAT6* gene via a complex regulatory network.

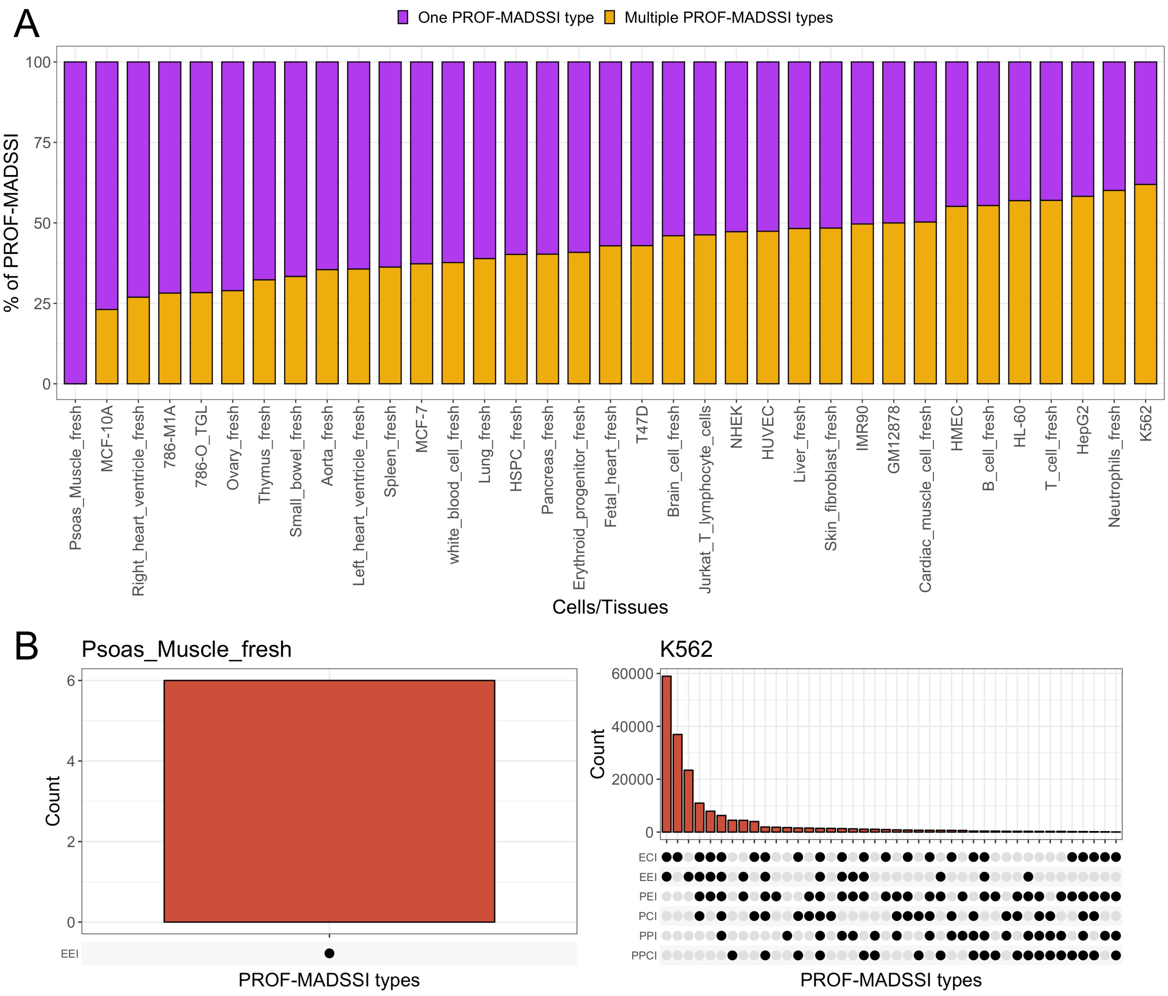


Figure 5: Investigation of PROF-MADSSI types in the identified PROF-MADSSI. (A) bar plot shows the percentage of PROF-MADSSI that are identified as only one PROF-MADSSI type or identified as multiple PROF-MADSSI types in each cell/tissue. (B) Upset plots show the frequency of each PROF-MADSSI type in two extreme examples, including psoas muscle (left), which has only one PROF-MADSSI types and K562 (right), which have the most (61.9%) of PROF-MADSSI are identified as multiple PROF-MADSSI types.

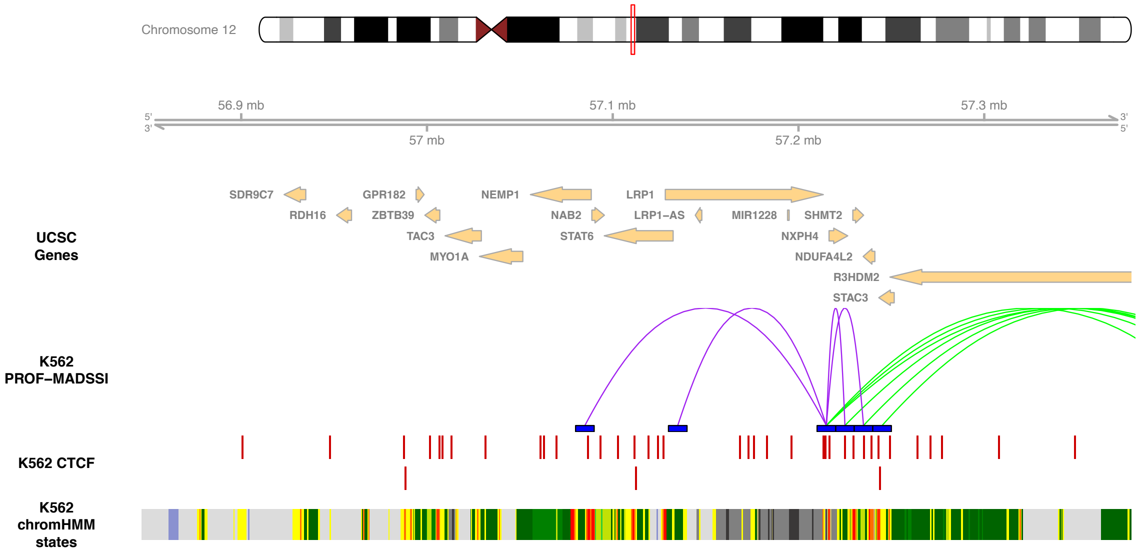


Figure 6: An example of a chromatin interaction being annotated as all six PROF-MADSSI types in K562 cell line. From the top, there are the UCSC gene track shows the known genes in the plotting region (chr12:56.9-57.3 Mb), interaction track shows the PROF-MADSSI of K562, annotation track shows the CTCF binding sites and chromHMM states track indicate the predicted K562-specific chromHMM states.

We then investigated the genomic distance distribution of each PROF-MADSSI type across 35 cell/tissue types and found that the vast majority of identified PROF-MADSSI were located within 2 Mb (Supplementary Figure 4), relatively shorter than previous 10 Mb length of previous informative interactions estimates [32,42]. Furthermore, there were two types of distribution patterns observed across 35 cells/tissues: one peaks at around 200-250 kb, followed by a decay curve, such as in B cells (Figure 7), Brain cells, Cardiac Muscle cells etc (Supplementary Figure 4). The other pattern peaks at 10 kb (adjacent interacting bins), such as HepG2 (Figure 7), HL-60, lung etc (Supplementary Figure 4). Additionally, we found that the distributions of different PROF-MADSSI varied slightly in the first pattern whereas they are more consistent in the second pattern. For example, for the genomic distance of 200-250 kb in B cells (Figure 7), all PORF-MADSSI types are more frequently observed than all MADSSI, with EEI the most frequent PROF-MADSSI type and PPCI the least frequent. This indicates that the potential regulatory functional interactions are more likely to occur over short distances (200-250 kb) than other MADSSI in B cells. However, in HepG2 (Figure 7), the distributions of all PROF-MADSSI types are similar to all MADSSI.

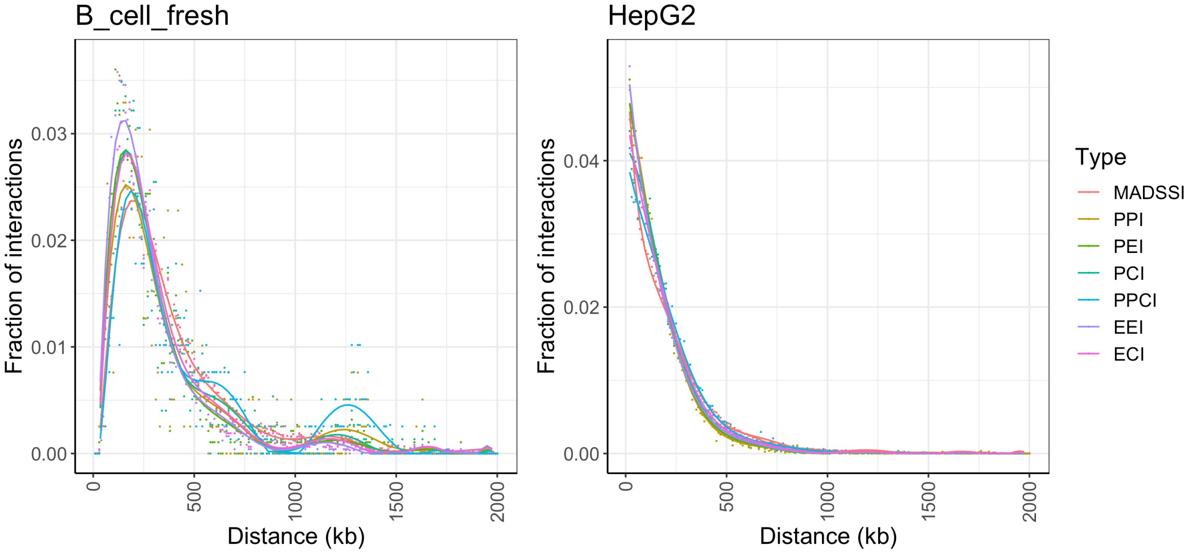


Figure 7: Investigation of the distance distribution between six PROF-MADSSI types. Line plots show the genomic distance distribution of MADSSI and PROF-MADSSI in B cells (left) and HepG2 (right).

Identification of 3D regulatory regions from PROF-MADSSI

Given the complex nature of the gene regulation network, it has always been challenging to link regulatory elements to their target genes [32]. This particular problem is essential to the analysis of targets identified from Genome-Wide Association Studies (GWAS), where historically, genes were assigned to lead SNPs based on linear genomic distances (i.e. the closest gene). Using cell/tissue type-specific PROF-MADSSI that contacting with cell/tissue-specific expressed gene promoters directly, including PPI, PEI, PCI and PPCI, we generated lists of cell/tissue-specific 3D regulatory regions (i.e. the corresponding promoters, enhancers and CTCF binding sites) that are located in close proximity to cell/tissue-specific expressed gene promoters in 3D space, with the potential to mediate gene regulation via physical contact (Figure 8A and Supplementary File 3). In addition to the direct linkage between regulatory regions to expressed gene promoters, we also considered secondary linkages, which are mainly driven by EEI and ECI (Figure 8A), hence identifying extra 3D regulatory regions (i.e. enhancers with secondary linkages). Overall, in 35 cells/tissues, only enhancers-related PROF-MADSSI (ECI and EEI) were detected in Psoas Muscle and Small Bowel tissue, they were therefore excluded from the following analysis. Consistently to the finding of EEI and ECI made up of the vast majority of PROF-MADSSI, on average 66.98% of the 3D regulatory regions are composed of enhancers, including non-expressed active enhancers (60.6%), bivalent enhancers (4.86%) and expressed enhancers (1.52%), while CTCF binding sites (19%) being the second largest component (Figure 8B).

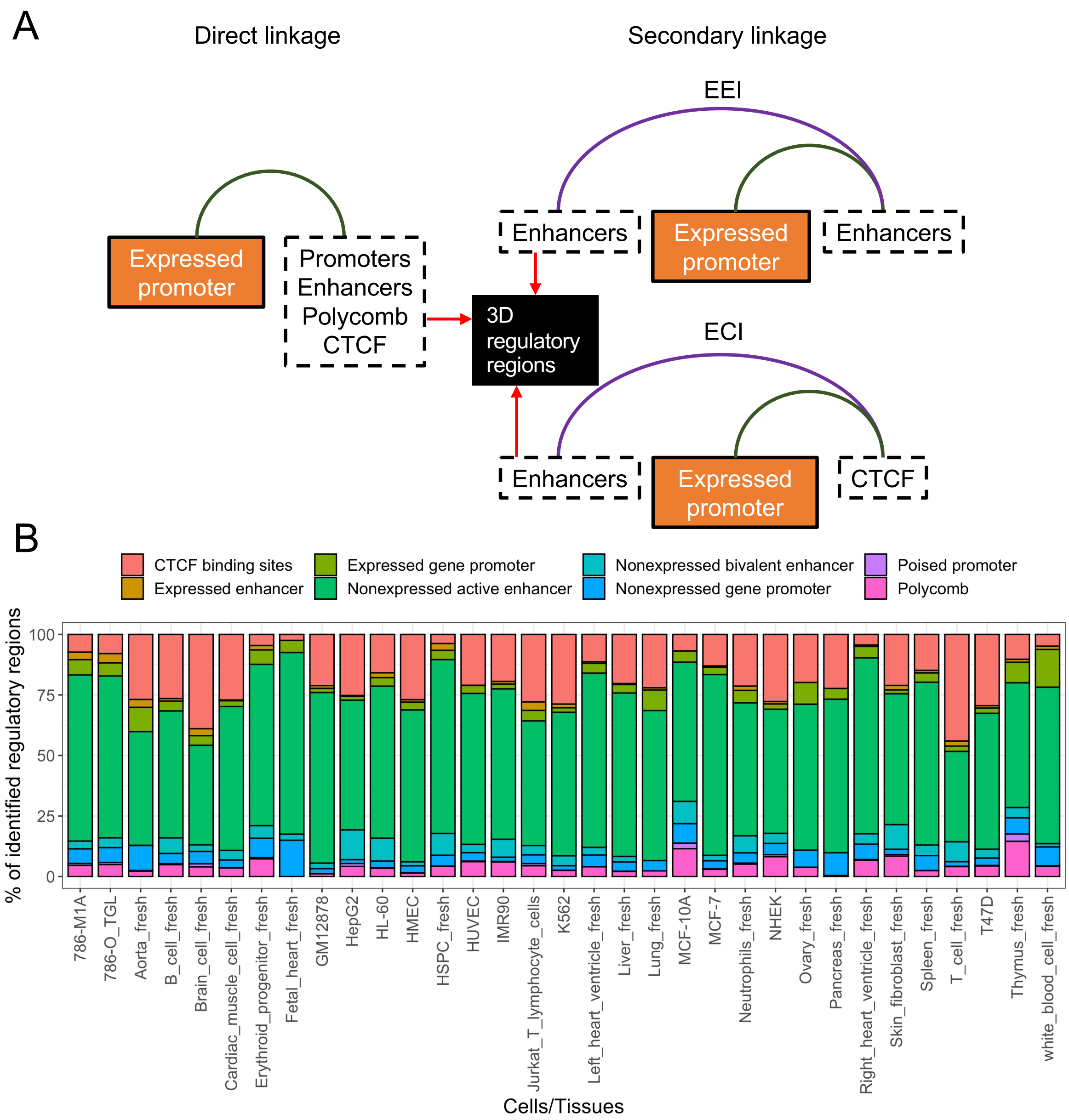


Figure 8: The definition and identification of 3D regulatory regions. (A) Schematic figure shows the direct linkage and secondary linkage that used to identify 3D regulatory regions in this study. (B) Bar plot shows the composition of 3D regulatory regions in each cell/tissue.

To demonstrate the potential regulatory functionality of 3D regulatory regions, we further integrated the 3D regulatory regions of each cell/tissue type with cell/tissue-specific super-enhancers (SEs), which play an important role in transcription factor binding and gene expression regulation in the genome [43,44]. We obtained from the Super-Enhancer Archive (SEA) [45], finding 68% (15 out of 22) of the cell/tissue-specific SEs are enriched in their corresponding type-specific SEs (Figure 9). For instance, T cell, neutrophils, MCF-7 and liver-specific SEs were enriched in the 3D regulatory regions of T cells, neutrophils, MCF-7 and liver tissue, respectively. The exceptions being 3D regulatory regions of thymus, spleen, lung, aorta, pancreas, ovary and right heart ventricle tissues, all of which failed to show enrichment in their corresponding tissue-specific SEs (Figure 9). We also observed an overall high enrichment between 3D regulatory regions of IMR90, GM12878 and K562 and all cell/tissue-specific SEs (Figure 9). By clustering of the similar enrichment pattern of SEs, we found some clustering patterns associated with the cell/tissue type specificity, such as T cells and Jurkat T cell lines are clustered together with a similar enrichment pattern across all SEs and exhibit high enrichment level in T cell-specific SEs, and other blood-related cell types, including HSPC, white blood cells, B cells, neutrophils and erythroid progenitor cells are clustered together (Figure 9).

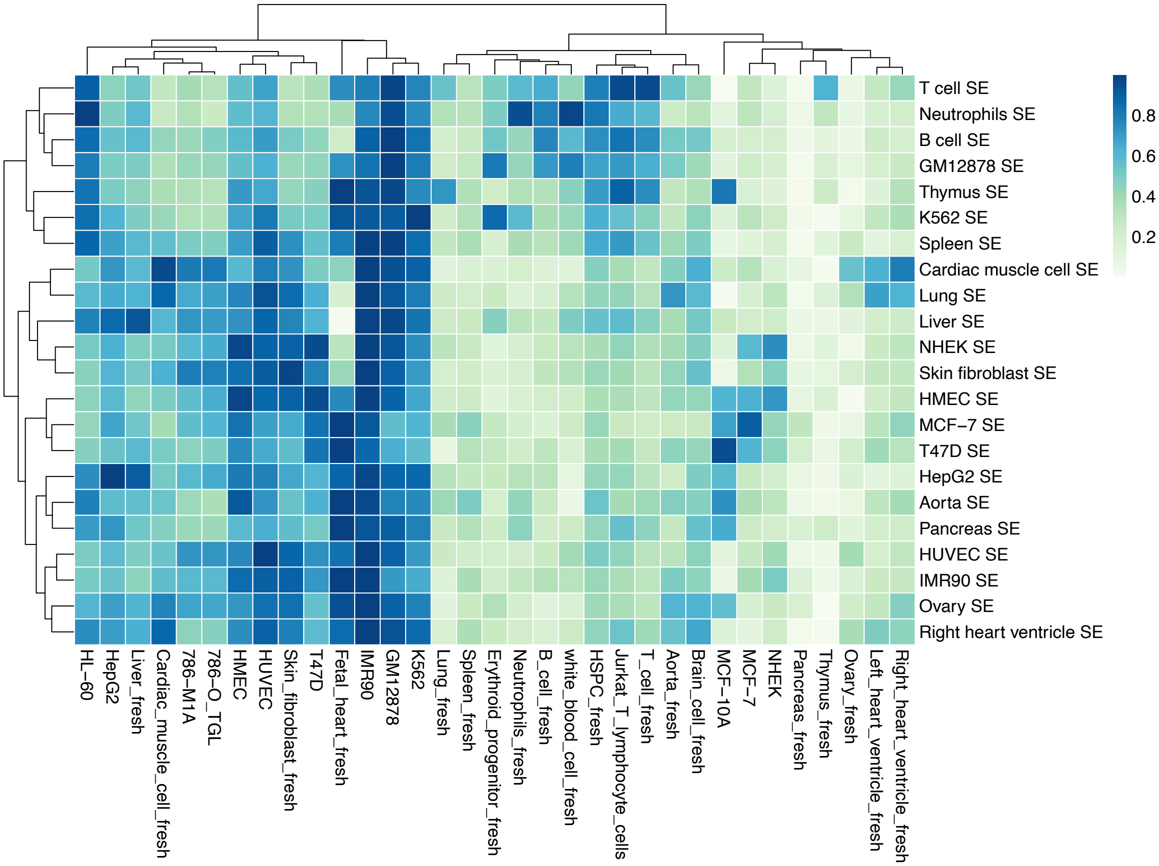


Figure 9: Investigation of the enrichment of SEs in 3D regulatory regions. Heatmap shows the enrichment level of cell/tissue-specific super-enhancers in the identified cell/tissue-specific 3D regulatory regions.

We finally integrated 3D regulatory regions with tissue-specific *cis*- (distance between eQTLs and their target genes within 1 Mb) expression quantitative trait loci (eQTLs) obtained from the GTEx database [46]. In 15 cells/tissues with corresponding eQTLs, eQTLs could be detected in from 11% (liver) to 43% (Erythroid progenitor) of the identified cell/tissue-specific 3D regulatory regions (Figure 10A). By comparing the 3D regulatory region-gene promoter interactions to eQTL-target gene pairs, we found on average 5.9% of the 3D regulatory region-gene promoter interactions overlapped with the eQTL-target gene pairs in the corresponding cells/tissues (Figure 10B), meaning these eQTLs are located within a 3D regulatory region (i.e. CTCF binding sites, enhancers or promoters) and having physical contact via 3D chromatin interactions with the promoters of their target genes, whose expression levels are altered by the presence of the eQTLs in the genome. Here we characterise these as “3DeQTLs” and we identified cell/tissue-specific 3DeQTLs in 15 cells/tissues (Supplementary File 4). Even though only a small fraction (on average 0.29%) of eQTLs were identified as 3DeQTLs, on average 1.15%, 8.5% and 16.1% of the eQTLs were captured by cell/tissue-specific 3D regulatory regions, PROF-MADSSI and MADSSI, respectively (Figure 10C). Some cells/tissues-specific MADSSI, such as skin fibroblast, brain cells and T cells, were found to be overlapped over 40% of the eQTLs, indicating the regulatory importance of the cataloged MADSSI (Figure 10C).

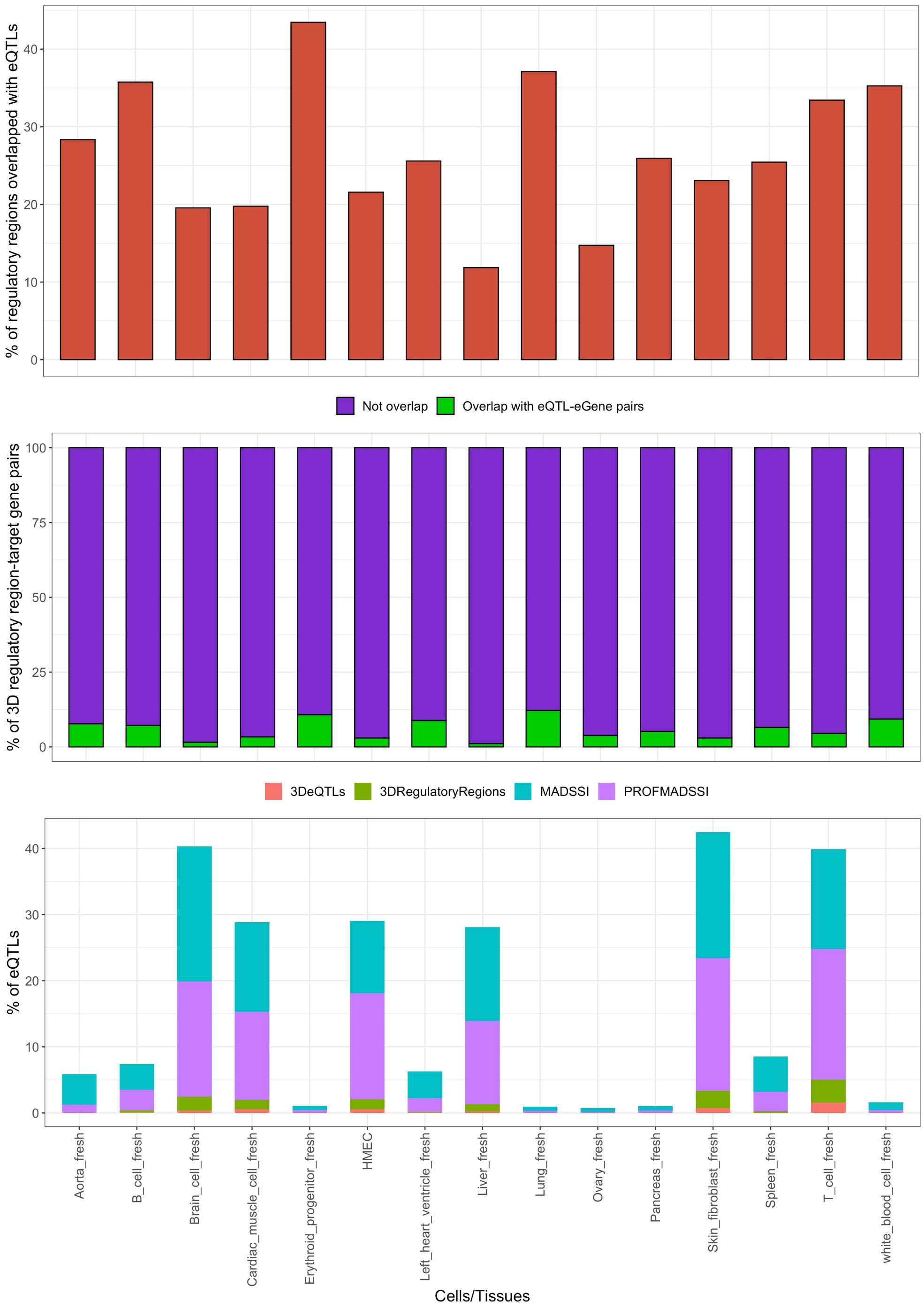


Figure 10: Investigation of eQTLs integrating with 3D regulatory regions. (Top panel) bar plot shows the percentage of 3D regulatory regions that overlapped with cell/tissue-specific *cis* eQTLs. eQTLs are obtained from the GTEx database [46]. (Middle panel) bar plot shows percentage of 3D regulatory region-gene promoter interactions overlap with eQTL-target gene pairs. (Bottom panel) bar plot shows the percentage of *cis* eQTLs found in MADSSI, PROF-MADSSI, 3D regulatory regions and identified as 3DeQTLs.

In order to demonstrate the relationship between 3D regulatory regions and the 3DeQTLs we identified, we used the 3DeQTLs found in the liver tissue. Secreted phosphoprotein-1 (*SSP1*) gene, found on chromosome 4, is responsible for the production of osteopontin and was found to play an important role in the development of liver diseases and hepatocellular carcinoma [47–49]. More importantly, the *SSP1* gene is also identified as an eQTL target gene for many liver-specific *cis* eQTLs located in the non-coding region upstream of the *SSP1* gene (Figure 11). Liver-specific MADSSI were observed between the promoter region (2 kb upstream of the transcription start site) of *SSP1* gene and liver-specific 3D regulatory regions (shown as orange) (Figure 11). These 3D regulatory regions contain 4 liver-specific CTCF-binding sites and 2 active enhancer chromHMM states, and surrounded by 149 eQTLs targeting *SSP1* located between the CTCF binding sites of one 3D regulatory region (on the left) and *SSP1* promoter, indicating that this is a CTCF-mediated enhancer-promoter chromatin loop linking eQTLs within the loop to contact their target gene promoter (Figure 11). Interestingly, we found that this loop is located near the boundary of a topologically-associated domain (TAD), and the interaction between the 3DeQTLs and *SSP1* promoter is a cross-TAD interaction (Figure 11). Altogether, this highlights the potential mechanism of eQTLs mediating the expression of their target genes by associating with the 3D regulatory regions which contact the promoters of target genes three-dimensionally.

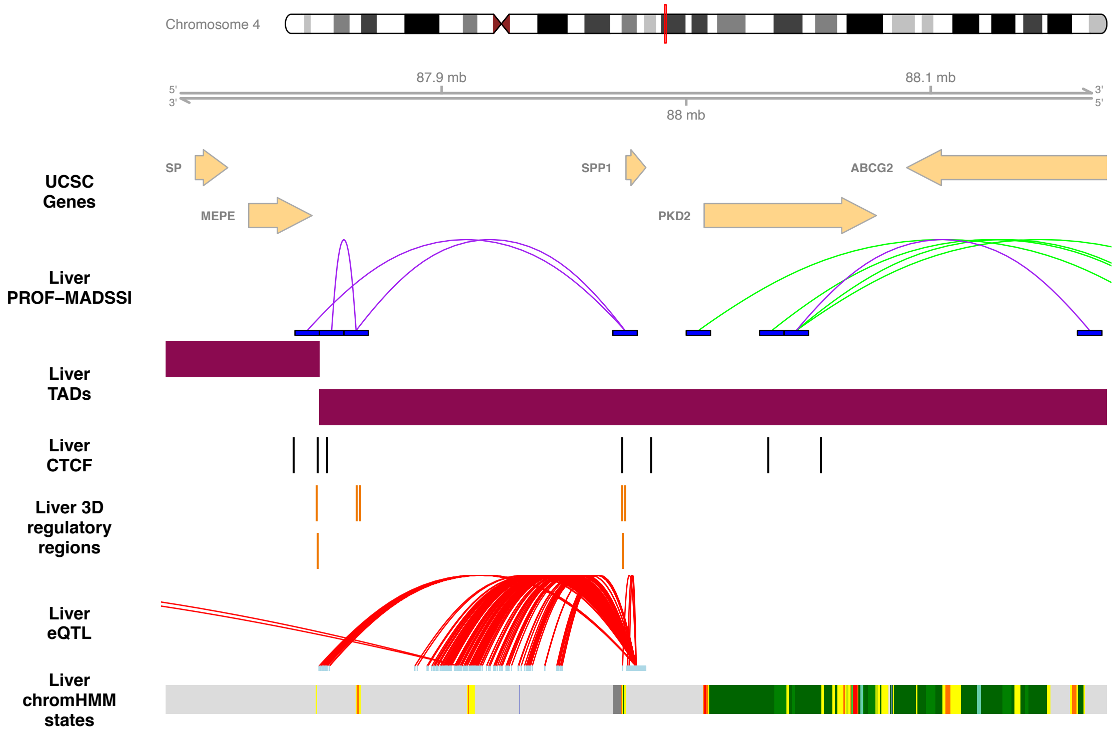


Figure 11: An example demonstrating the potential mechanism of 3DeQTL and its target gene in liver tissue. From the top, there are the UCSC gene track shows the known genes in the plotting region (chr4:87.9-88.1 Mb), interaction track shows the liver PROF-MADSSI, annotation tracks indicate liver TADs, CTCF binding sites and 3D regulatory regions, interaction track shows the linkage between liver *cis* eQTLs and *SSP1* gene and chromHMM states track indicate the predicted liver-specific chromHMM states.

Discussion

Chromatin interactions have been proved to play important roles in the gene regulation networks [2,6,50–53]. It is essential to accurately distinguish likely functional interactions from Hi-C datasets while accounting for biases such as random ligations, self ligations and uneven digestion patterns. To address this issue, using the statistical background model from MaxHiC [14] and publicly available Hi-C datasets, we generated MaxHiC-detected statistically significant interaction (MADSSI) profiles for 51 cell lines and tissues in previous works. In this study, to comprehensively annotate the cell/tissue-specific MADSSI profiles, we categorised chromatin interactions into 66 interaction classes based on the overlap between cell/tissue-specific annotations and each anchor bin of an interaction.

From 66 interaction types, based on different 3D regulatory models, which associate with promoters, enhancers, CTCF binding sites and polycomb complex, that have been studied [29,30,32,34,37,40], we selected relevant interaction types as potentially regulatory functional MADSSI (PROF-MADSSI) (Table 1). Using the identified PROF-MADSSI, we can easily reveal regions with regulatory signals, such as the STAT6 region we illustrated in K562 (Figure 6). As an important transcription factor, the presence of phosphorylated STAT6 is shown to play important role in the activation of the janus kinase-signal transducer and activator of transcription signaling pathway, which is essential for lymphoma and leukemia [41,54–56]. Our identification of the novel linkage (PROF-MADSSI) between the promoter of *STAT6* and the distal (100 kb upstream) regions, where CTCF binding sites, polycomb complex and enhancer signals are found, indicate an CTCF-drive promoter-enhancer/polycomb interactions may be involved in the regulation of expression of *STAT6* in K562, a chronic myelogenous leukemia-derived cell line. This potential 3D regulation can further be examined by knockout experiments via genome editing techniques such as CRISPR.

Different studies used different overlapping thresholds to determine successful annotations of chromatin interactions, some used an at least 10% overlap as a threshold [57], and some decided that any overlap is sufficient [32]. In our study, in order to minimise false positive annotations, we required a 100% overlap between annotations and interacting bins. However, this in turn might lead to an uneven distribution of the detected interaction types (Figure 1B), where annotations with smaller size, such as active enhancers that are defined by chromHMM states, are more likely to be annotated.

Spatial transcriptomics studies have recently shown the variation of gene expression profiles across different parts of a tissue and demonstrated the importance of positional context of gene expression in understanding the tissue functionality and its disease-associated pathological change [58–60]. This specificity is crucially important in defining epigenomic annotations and chromatin states, with heterogeneous tissues made up of multiple defined cell-types potentially leading to inaccurate results. Since chromatin interactions contribute to govern the gene expression regulations, it is reasonable to hypothesise that different parts of tissue exhibit different genome structure profiles. In this study, we used annotations from large epigenomic projects such as ENCODE [26] and the Roadmap Epigenomics [61] to annotate the tissue-specific MADSSI, but we were forced to neglect the spatial context across the tissue. In the future, integrating spatial transcriptomics data with spatial chromatin interactions may help us to understand the mechanism governed by 3D genome structure more comprehensively and accurately.

The functional mechanism of how non-coding variants, such as eQTLs affect gene expression can be interpreted by 3D regulatory regions and MADSSI. By integrating with cell/tissue-specific *cis* eQTLs, we showed a small fraction (0.29%) of eQTLs can be classified as 3DeQTLs, which are located in 3D regulatory regions and contact to the promoter region of their target genes, demonstrating the power of using PROF-MADSSI to functionally interpret eQTLs. Further validations can be carried out by regulatory region-knockout experiments in parallel with SNP editing experiments using CRISPR gene editing to validate the role of chromatin interactions in the functional mechanism of eQTLs. Future research of eQTLs and other non-coding SNPs interpretation can be facilitated by the cell/tissue-specific 3D regulatory regions and PROF-MADSSI catalogued in this study.

Conclusion

In conclusion, we comprehensively annotated cell/tissue-specific MADSSI for 35 human cells and tissue, generating lists of interactions and regions that have potential regulatory functions in a cell/tissue-specific manner. These results can further be used in cell/tissue type-specific data integration studies, such as identifying potential GWAS SNPs in cell or tissue types specifically associated with disease [3,62] or identifying novel non-coding regulatory regions such as long-non coding RNAs [63–65]. In future studies, based on more and more publicly available clinical Hi-C datasets, we will be able to generate disease-specific MADSSI profiles from clinical samples and compare them with our reference MADSSI profiles to further improve the interpretation of the regulatory function of disease/cell/tissue-specific chromatin interactions and regulatory regions.

Methods

Interaction classes annotation

In order to stringently annotate cell/tissue-specific MADSSI into various interaction classes, we required 100% overlap between annotation data and interacting bins for them to be defined as successfully annotated for each identified MADSSI. Overlaps between the annotation and interaction was carried out using the *pybedtools* library [66,67].

Enrichment test of super-enhancers

In order to perform enrichment analysis between 3D regulatory regions and super-enhancers, we first created a background of 3D regulatory regions by merging all 3D regulatory regions of all cells/tissues. For each cell/tissue and each type-specific super-enhancers, we then calculated the expected overlap count of super-enhancers by permuting the same amount of regions as the 3D regulatory regions 10 times from the background, the mean of the resulting overlapped count then being used as the expected overlap count. The enrichment score is then calculated as the observed overlap count divided by the expected overlap count for each cell/tissue-specific 3D regulatory region and super-enhancer. To visualise the enrichment, for each cell/tissue-specific enhancer, the enrichment scores underwent z score normalisation, followed by calculating the percentile rank using the *pnorm* function from the *stats* R library.

Interaction and eQTL visualisation

To visualise the identified interactions, 3D regulatory regions and eQTL linkage in specific regions such as Figure 3F, we used functions from R package *Gviz [68]*, *GenomicInteractions* [69] and *rtracklayer* [70].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

See supplementary documents.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

NL and JB developed and conceived the study in collaboration with HAR. NL wrote and developed approximately 90% of the manuscript, with editing provided by HAR and JB.

Acknowledgements

We acknowledged the publicly available Hi-C datasets and epigenome data.

References

1. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013;503:290–4.

2. 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;167:1369–84.e19.

3. Greenwald WW, Chiou J, Yan J, Qiu Y, Dai N, Wang A, et al. Pancreatic islet chromatin accessibility and conformation defines distal enhancer networks of type 2 diabetes risk. BioRxiv. 2018;299388.

4. 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;326:289–93.

5. 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;485:376–80.

6. Rao SSP, 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;159:1665–80.

7. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep. 2016;17:2042–59.

8. 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;47:598–606.

9. Ogiyama Y, Schuettengruber B, Papadopoulos GL, Chang J-M, Cavalli G. Polycomb-Dependent Chromatin Looping Contributes to Gene Silencing during Drosophila Development. Mol Cell. 2018;71:73–88.e5.

10. Ay F, Bailey TL, Noble W. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. 2014;24:999–1011.

11. 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:e0174744.

12. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012;28:3131–3.

13. Carty M, Zamparo L, Sahin M, González A, Pelossof R, Elemento O, et al. An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. Nat Commun. 2017;8:ncomms15454.

14. Alinejad-Rokny H, Ghavami R, Rabiee HR, Rezaei N. MaxHiC: robust estimation of chromatin interaction frequency in Hi-C and capture Hi-C experiments. bioRxiv [Internet]. biorxiv.org; 2020; Available from: https://www.biorxiv.org/content/10.1101/2020.04.23.056226v1.abstract

15. Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, et al. The 4D nucleome project. Nature. 2017;549:219–26.

16. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, et al. The European Nucleotide Archive. Nucleic Acids Res. 2011;39:D28–31.

17. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489:109–13.

18. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012;148:84–98.

19. Beagrie RA, Scialdone A, Schueler M, Kraemer DCA, Chotalia M, Xie SQ, et al. Complex multi-enhancer contacts captured by genome architecture mapping. Nature. 2017;543:519–24.

20. Ron G, Globerson Y, Moran D, Kaplan T. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nat Commun. 2017;8:2237.

21. Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, et al. Expression Atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Res. academic.oup.com; 2018;46:D246–51.

22. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47:D766–73.

23. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.

24. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

25. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017;12:2478–92.

26. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

27. Ziebarth JD, Bhattacharya A, Cui Y. CTCFBSDB 2.0: a database for CTCF-binding

sites and genome organization. Nucleic Acids Res. 2013;41:D188–94.

28. Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre B-M, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. 2015;25:582–97.

29. Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature. 2016;539:452–5.

30. Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, et al. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat Methods. 2017;14:629–35.

31. 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;51:1442–9.

32. Chen H, Xiao J, Shao T, Wang L, Bai J, Lin X, et al. Landscape of Enhancer-Enhancer Cooperative Regulation during Human Cardiac Commitment. Mol Ther Nucleic Acids. 2019;17:840–51.

33. Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell Res. 2012;22:490–503.

34. Ing-Simmons E, Seitan VC, Faure AJ, Flicek P, Carroll T, Dekker J, et al. Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. Genome Res. 2015;25:504–13.

35. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. 2016;15:2038–49.

36. Davidson IF, Bauer B, Goetz D, Tang W, Wutz G, Peters J-M. DNA loop extrusion by human cohesin. Science. 2019;366:1338–45.

37. Kubo N, Ishii H, Xiong X, Bianco S, Meitinger F, Hu R, et al. Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. Nat Struct Mol Biol. 2021;28:152–61.

38. Aranda S, Mas G, Di Croce L. Regulation of gene transcription by Polycomb proteins. Sci Adv. 2015;1:e1500737.

39. Schoenfelder S, Sugar R, Dimond A, Javierre B-M, Armstrong H, Mifsud B, et al. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nat Genet. 2015;47:1179–86.

40. Kar G, Kim J, Kolodziejczyk AA, Natarajan K, Triglia E, Mifsud B, et al. Flipping between Polycomb repressed and active transcriptional states introduces noise in gene expression. Nat Commun. 2017;8:36.

41. Bruns HA, Kaplan MH. The role of constitutively active Stat6 in leukemia and lymphoma. Crit Rev Oncol Hematol. 2006;57:245–53.

42. Sahlén P, Abdullayev I, Ramsköld D, Matskova L, Rilakovic N, Lötstedt B, et al. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. Genome Biol. 2015;16:156.

43. He Y, Long W, Liu Q. Targeting Super-Enhancers as a Therapeutic Strategy for Cancer Treatment. Front Pharmacol. 2019;10:361.

44. Lee B-K, Jang YJ, Kim M, LeBlanc L, Rhee C, Lee J, et al. Super-enhancer-guided mapping of regulatory networks controlling mouse trophoblast stem cells. Nat Commun. 2019;10:4749.

45. Chen C, Zhou D, Gu Y, Wang C, Zhang M, Lin X, et al. SEA version 3.0: a comprehensive extension and update of the Super-Enhancer archive. Nucleic Acids Res. 2020;48:D198–203.

46. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.

47. Lorena D, Darby IA, Gadeau A-P, Leen LLS, Rittling S, Porto LC, et al. Osteopontin expression in normal and fibrotic liver. altered liver healing in osteopontin-deficient mice. J Hepatol. 2006;44:383–90.

48. Shin HD, Park BL, Cheong HS, Yoon J-H, Kim YJ, Lee H-S. SPP1 polymorphisms associated with HBV clearance and HCC occurrence. Int J Epidemiol. 2007;36:1001–8.

49. Wen Y, Jeong S, Xia Q, Kong X. Role of Osteopontin in Liver Diseases. Int J Biol Sci. 2016;12:1121–8.

50. Smith E, Shilatifard A. Enhancer biology and enhanceropathies. Nat Struct Mol Biol. 2014;21:210–9.

51. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science. 2016;351:1454–8.

52. Rickels R, Shilatifard A. Enhancer Logic and Mechanics in Development and Disease. Trends Cell Biol. 2018;28:608–30.

53. Di Giammartino DC, Polyzos A, Apostolou E. Transcription factors: building hubs in the 3D space. Cell Cycle. 2020;19:2395–410.

54. Guiter C, Dusanter-Fourt I, Copie-Bergman C, Boulland M-L, Le Gouvello S, Gaulard P, et al. Constitutive STAT6 activation in primary mediastinal large B-cell lymphoma. Blood. 2004;104:543–9.

55. Ritz O, Guiter C, Castellano F, Dorsch K, Melzner J, Jais J-P, et al. Recurrent mutations of the STAT6 DNA binding domain in primary mediastinal B-cell lymphoma. Blood. 2009;114:1236–42.

56. Kaymaz BT, Selvi N, Gündüz C, Aktan Ç, Dalmızrak A, Saydam G, et al. Repression of STAT3, STAT5A, and STAT5B expressions in chronic myelogenous leukemia cell line

K–562 with unmodified or chemically modified siRNAs and induction of apoptosis [Internet]. Annals of Hematology. 2013. p. 151–62. Available from: http://dx.doi.org/10.1007/s00277-012-1575-2

57. Qin Y, Grimm SA, Roberts JD, Chrysovergis K, Wade PA. Alterations in promoter interaction landscape and transcriptional network underlying metabolic adaptation to diet. Nat Commun. 2020;11:962.

58. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353:78–82.

59. Asp M, Salmén F, Ståhl PL, Vickovic S, Felldin U, Löfling M, et al. Spatial detection of fetal marker genes expressed at low level in adult human heart tissue. Sci Rep. 2017;7:12941.

60. Yoosuf N, Navarro JF, Salmén F, Ståhl PL, Daub CO. Identification and transfer of spatial transcriptomics signatures for cancer diagnosis. Breast Cancer Res. 2020;22:6.

61. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. nature.com; 2010;28:1045–8.

62. Liu N, Sadlon T, Wong YY, Pederson S, Breen J, Barry SC. 3DFAACTS-SNP: Using regulatory T cell-specific epigenomics data to uncover candidate mechanisms of Type-1 Diabetes (T1D) risk [Internet]. Cold Spring Harbor Laboratory. 2020 [cited 2020 Nov 25]. p. 2020.09.04.279554. Available from: https://www.biorxiv.org/content/10.1101/2020.09.04.279554v1.abstract

63. Hou Y, Zhang R, Sun X. Enhancer lncRNAs influence chromatin interactions in different ways. Front Genet. 2019;10:936.

64. Agrawal S, Alam T, Koido M, Kulakovskiy IV, Severin J. Functional annotation of human long noncoding RNAs using chromatin conformation data. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.01.13.426305v1.abstract

65. Statello L, Guo C-J, Chen L-L, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol. 2021;22:96–118.

66. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

67. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics. 2011;27:3423–4.

68. Hahne F, Ivanek R. Visualizing Genomic Data Using Gviz and Bioconductor. Methods Mol Biol. 2016;1418:335–51.

69. Harmston N, Ing-Simmons E, Perry M, Barešić A, Lenhard B. GenomicInteractions: An R/Bioconductor package for manipulating and investigating chromatin interaction data. BMC Genomics. 2015;16:963.

70. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25:1841–2.

Supplementary information

Note: Supplementary Tables and Files are hosted on Figshare:

Supplementary Table 1: https://figshare.com/s/36e06d866750857fc67c

Supplementary File 1: https://figshare.com/s/782e887851131456fc68

Supplementary File 2: https://figshare.com/s/138a2d4fccf0a35ca64c

Supplementary File 3: https://figshare.com/s/b391172df823df74ec4f

Supplementary File 4: https://figshare.com/s/b5c567350499f33e569f

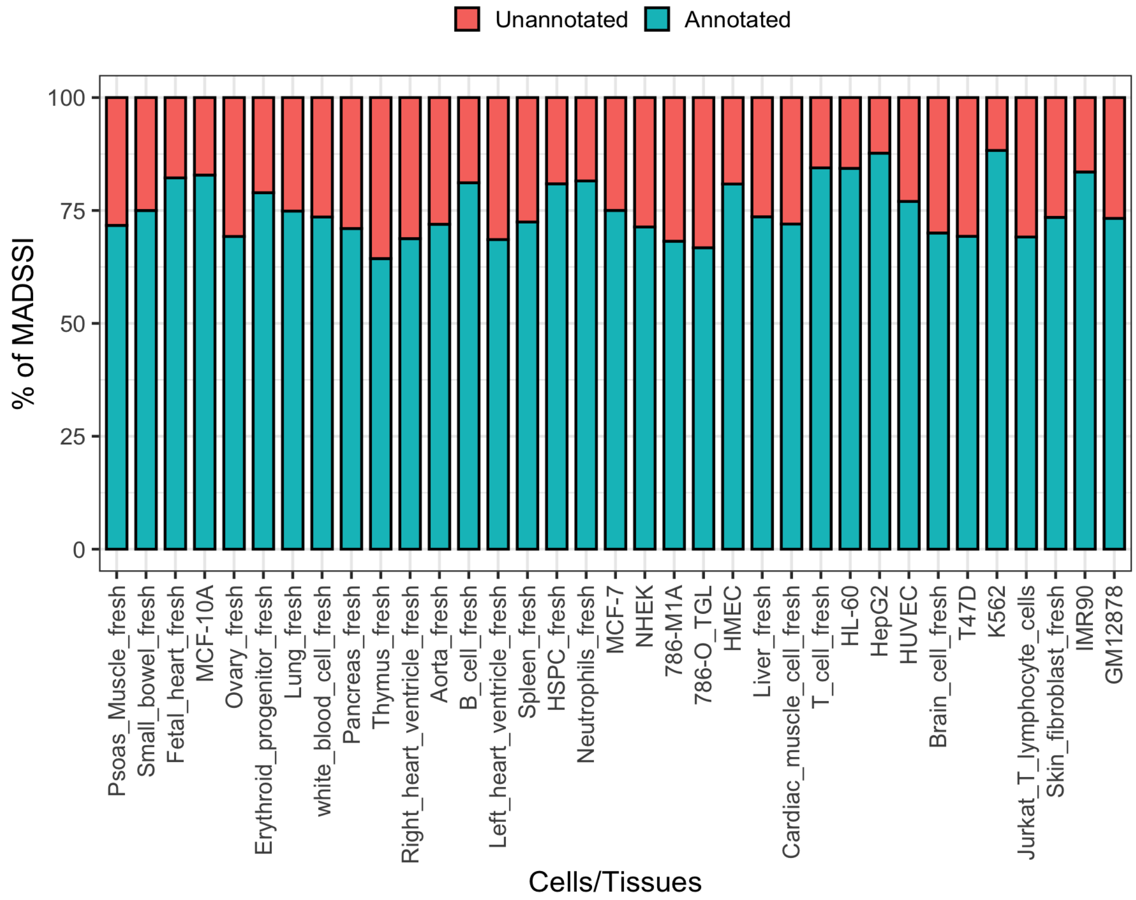
Supplementary Table 1: Cell/tissue-specific annotations used to annotate MADSSI in this study.

Supplementary File 1: Annotated MADSSI with 66 interaction classes of 35 cells and tissues.

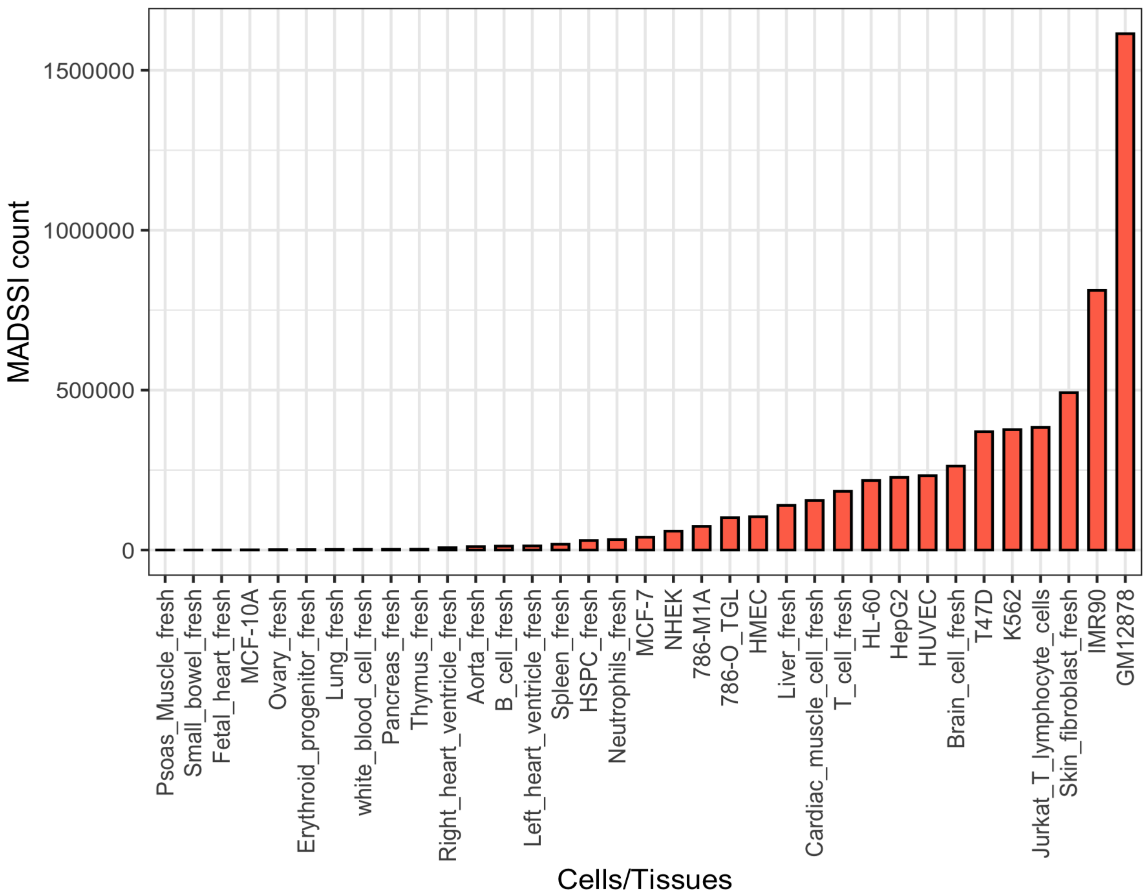
Supplementary File 2: PROF-MADSSI of 35 cells and tissues identified in this study.

Supplementary File 3: 3D regulatory regions of 33 cells and tissues identified in this study.

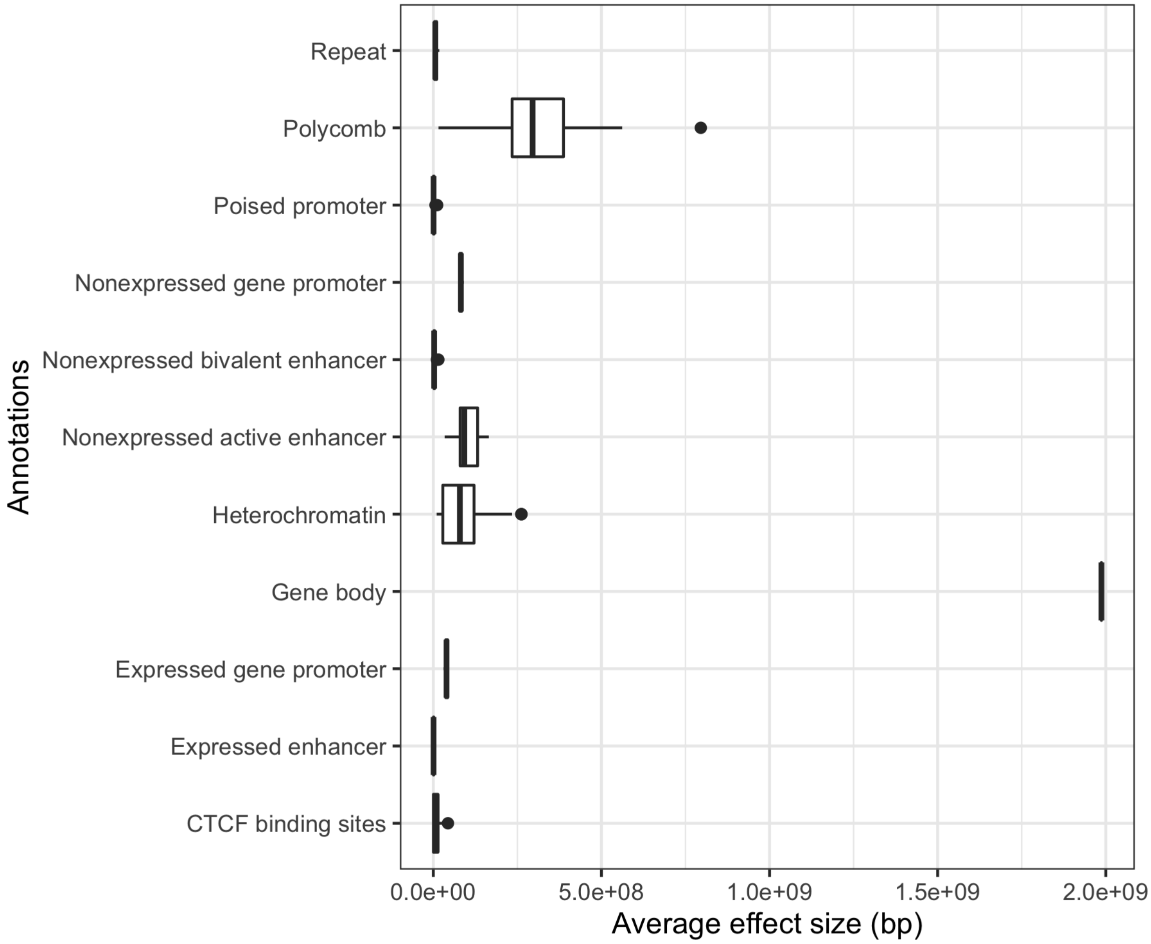
Supplementary File 4: 3DeQTLs of 15 cells and tissues identified in this study.



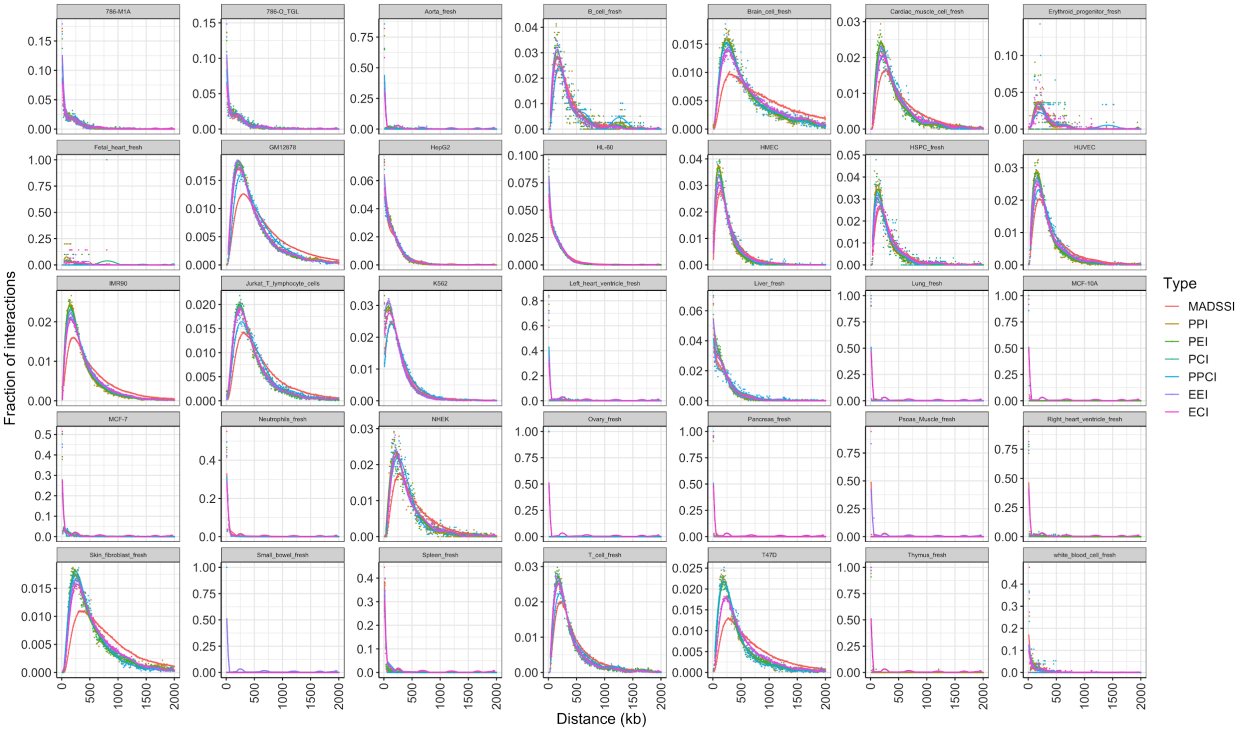
Supplementary Figure 1: The percentage of annotated interactions and unannotated interactions of all MADSSI of 35 cells/tissues.



Supplementary Figure 2: The MADSSI count of 35 cells/tissues.



Supplementary Figure 3: Average effect size of the 11 types of annotations used in this study. The average effect size is calculated by the product of average length and total count of each annotation in 35 cells and tissues.



Supplementary Figure 4: genomic distance distribution of PROF-MADSSI and MADSSI across 35 cell/tissue types.

Chapter 6

Conclusion and Discussion

Conclusion and Discussion

Over the last decade, a number of large epigenomics projects have been initiated to catalogue human cell and tissue type-specifics information, such as the Encyclopedia of DNA elements (ENCODE) project (ENCODE Project Consortium, 2012; Davis *et al.*, 2018), NIH Roadmap Epigenomics project (Bernstein *et al.*, 2010; Roadmap Epigenomics Consortium *et al.*, 2015), Functional Annotation of the Mammalian Genome (FANTOM5) project (Lizio *et al.*, 2015, 2019) and the Genotype-Tissue Expression (GTEx) project (Lonsdale *et al.*, 2013). These projects offer unprecedented access to functional information of individual cell-types across the genome, enabling researchers to investigate complex regulatory networks that may have a significant impact on common/complex disease and phenotypes that are unlikely to be impacted by simple variant systems. However, there are still limitations to these repositories, partly because of the highly cell type-specific regulation of gene expression in some compartments, especially the immune system, and because the accurate functional annotation of the linkage between non-gene-coding regions and genes requires newer conformation-specific datasets, which have not yet been generated for many cell types.

This means that despite the extensive, accurate annotation information of regulatory, non-coding regions (histone modifications, DNA methylation, enhancers, repressors etc), a direct functional connection to a specific gene target is often lacking, hindering the identification of causative gene regulation mechanisms that drive complex diseases. In this thesis, I investigated the dynamics of 3D chromatin interactions across the human genome, attempting to identify a tangible link between regulatory factors found in non-gene-coding regions and target genes via physical proximity in 3D space (Figure 1). This is important for correctly annotating genetic risk to the altered genes, and for filtering out SNPs in linkage disequilibrium that do not drive the alteration of gene expression because of the cell and condition-specific 3D chromatin organisation.

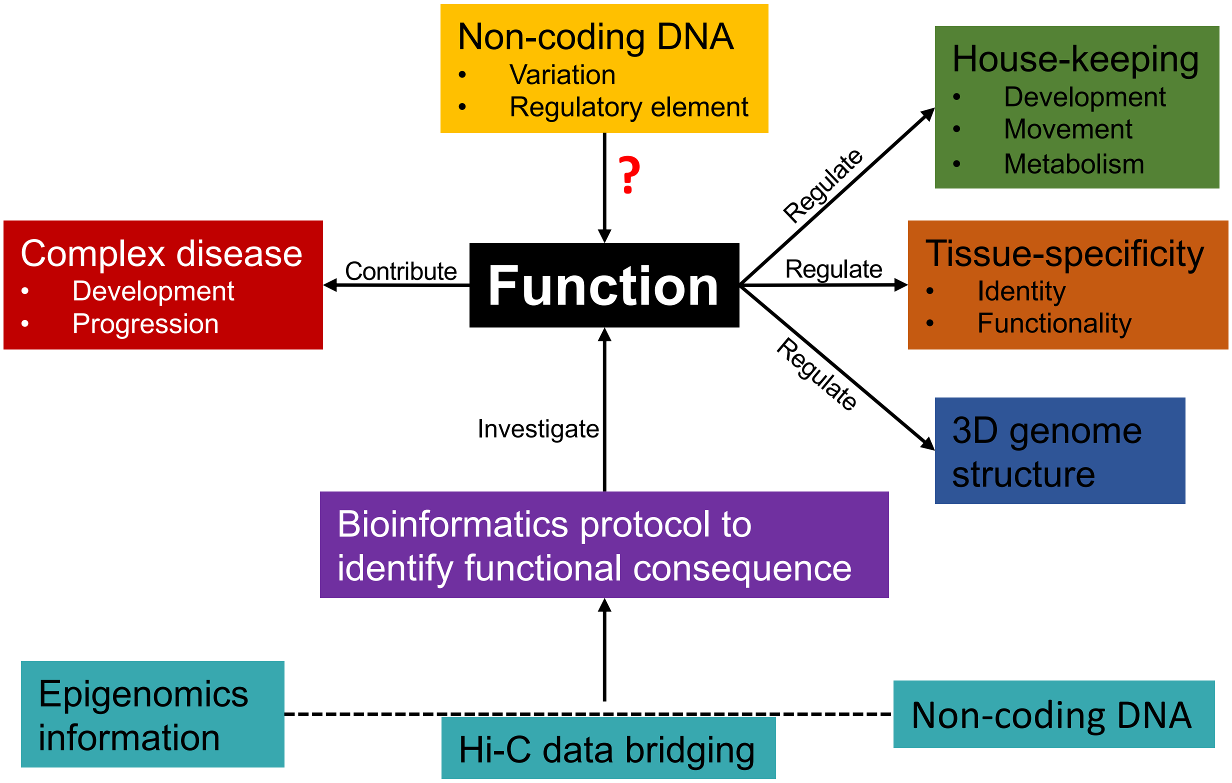


Figure 1: Research context of this PhD dissertation.

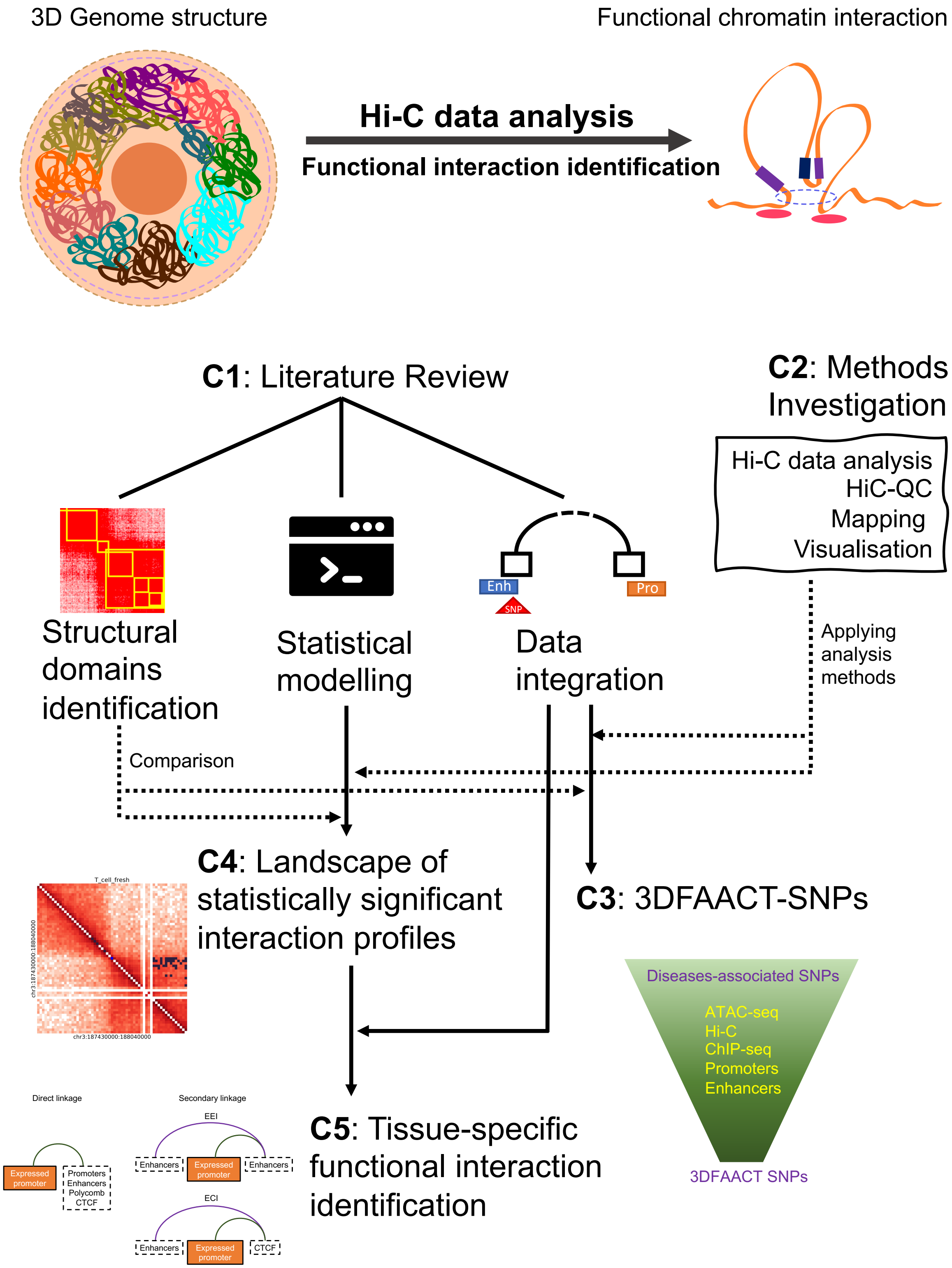


Figure 2: Schematic summary describing the relationship between each chapter of this dissertation.

Non-coding DNA: The dark matter of genetic risk

Classical case-control human genetics studies enabled researchers to discovered many important genetic alterations linked to diseases such as chronic granulomatous disease (Royer-Pokora *et al.*, 1986), cystic fibrosis (Kerem *et al.*, 1989) and Fanconi's anaemia (Strathdee *et al.*, 1992). Starting with few genetic markers that were commonly located in gene-coding regions, the development in high-throughput sequencing technologies and the reduction in the per-base cost of genome sequencing over the last 12 years has meant that researchers can now study the impact of genetic variation on a genome-wide scale. However, until very recently, genome-wide association studies (GWAS) were carried out using microarray technologies (Gorlov *et al.*, 2009; Liang *et al.*, 2020) that contained thousands of markers that were biased towards known, gene-rich regions, ignoring a wealth of non-coding variation. Despite this, the vast majority of the disease-associated single nucleotide polymorphisms (SNPs) discovered by GWAS have been located in non-coding regions (Freedman *et al.*, 2011; Tak and Farnham, 2015), with no link to genes other than by proximity. Even though many of the non-coding SNPs have been found located in regulatory regions, such as enhancers (Kikuchi *et al.*, 2019) and transcription factor binding sites (Huo *et al.*, 2019), it is still challenging to interpret their functionality without additional information of which genes they target.

Reported disease-associated variants from GWAS are commonly linked to target genes or regulatory elements by taking the closest gene by linear proximity

(Fernández and Miranda-Saavedra, 2012; International Multiple Sclerosis Genetics Consortium (IMSGC) *et al.*, 2013; Suna *et al.*, 2015; de Lange *et al.*, 2017). However, the genomic distance between the *bona fide* target genes and variants may be variable due to the nature of three-dimensional (3D) chromosome structure (Pennacchio *et al.*, 2013; Tak and Farnham, 2015; Hariprakash and Ferrari, 2019). Additionally, and crucial to the interpretation of GWAS results, target genes may not be associated with any genetic linkage pattern and could be found multiple Megabases (Mb) away from their regulatory partner. For example, Sanyal et al. 2012 demonstrated in three cell lines that only 47% of the regulatory elements interact with their nearest genes. 3D chromosome structure can bring distal elements such as enhancers or silencers and gene promoters in close physical proximity in 3D space, playing an important role in the gene regulation network (Lieberman-Aiden *et al.*, 2009; Jin *et al.*, 2013; Rao *et al.*, 2014; Mifsud *et al.*, 2015).

Given the important impact that the 3D chromosome structure has on physical proximity of genomic regions, 3D chromatin interactions can now be used to connect regulatory functionality to specific genes, revealing novel regulation mechanisms. These chromatin interactions can be captured through complex assays such as high-resolution chromosome conformation capture (Hi-C) sequencing assay (Lieberman-Aiden *et al.*, 2009), which allow detection of physical interactions within the nucleus by capturing crosslinked DNA using a digestion and re-ligation protocol. In the recent years, a significant amount of data has been generated across cell-lines, tissues and across species, all publicly available through databases such as the European nucleotide archive (ENA) (Leinonen *et al.*, 2011). These have largely been to facilitate the investigation of chromosome architecture, but they are now being used to interpret functional impacts of genetic studies. Once again, the functional impact of 3D connectivity has to be investigated in the relevant cell type to map causality of genetic risk of disease.

However, given the complexity of Hi-C datasets, it is difficult to interpret Hi-C data and identify chromatin interactions which drive biological functionality. One of the main reasons being the experimental procedures in the Hi-C protocol, particularly the ligation step, which can introduce false positive interactions to the Hi-C data due to scenarios such as self-ligation and random ligation events (Ay, Bailey and Noble, 2014; Mifsud *et al.*, 2017). Furthermore, Hi-C experiments are often performed with millions of cells and require high sequencing depth, such that the scale of data further complicates the identification of interactions with functional potential. Taken together, this suggests that in order to make use of the connectivity of Hi-C interactions to link non-coding DNA, such as non-coding SNPs, to their target genes, it is essential to accurately annotate real functional chromatin interactions from Hi-C datasets. In Chapter 1, we therefore reviewed the methodologies that are currently being used to identify potential functional interactions from Hi-C data (Figure 2). We categorised approaches of potentially functional interactions identification into three major groups: structural domain identification methods, including the detection of A/B compartments (Lieberman-Aiden *et al.*, 2009) and subcompartments (Rao *et al.*, 2014); detecting statistical significant interactions based on statistical modeling (Ay, Bailey and Noble, 2014; Carty *et al.*, 2017; Mifsud *et al.*, 2017; Alinejad-Rokny *et al.*, 2020) and data integration, including integrating with regulatory elements, genome-wide association studies (GWAS) and quantitative trait loci (QTL) data.

Using the reviewed methodologies to identify potential functional chromatin interactions, enable the discovery of 3D genome structure with interpretable biological functionality. We therefore can use them to establish linkages between non-coding DNA, specifically non-coding SNPs, and known genes, or gene regulators such as enhancers and repressors. We are then able to make predictions for novel genetic regulation mechanisms and identify previously unknown genetic risks for complex diseases, such as autoimmune diseases and cancers. These findings can potentially contribute to clinical genetic screens such as in the calculation of the polygenic risk scores, which is useful for predicting disease status and inherited susceptibility for individuals (Choi, Mak and O’Reilly, 2020; Lewis and Vassos, 2020), or assisting the development of novel genetic treatments for complex diseases such as enhancer therapies (Hamdan and Johnsen, 2019; Zhao Zhang *et al.*, 2019).

‘Not all SNPs are created equal’

As mentioned previously, a significant research gap lies in the interpretation of non-coding genetic variants identified from disease-specific GWAS studies. In Chapter 3, I address this research gap by building a computational pipeline called **3 D**imensional **F**unctional **A**nnotation of **A**ccessible **C**ell **T**ype-**S**pecific SNPs (3DFAACTS-SNP) to connect diseases-associated non-coding SNPs to their potential target genes based on a model of type 1 diabetes (T1D) regulatory mechanisms. As T1D is an autoimmune disease, regulatory T cells (Tregs), a cell-type that plays an important role in the homeostasis of the immune system, are implicated in the unrestrained immune destruction of the insulin-generating pancreatic beta cells (Atkinson, Eisenbarth and Michels, 2014), I hypothesised that T1D genetic risk in Treg cells would impact the development and progression of T1D. Therefore variants that specifically disrupt the regulatory mechanisms of Treg cells are perfect therapeutic targets for investigation. Based on the integration of T1D-associated SNPs with other genomic datasets, I developed a T1D-specific 3DFAACT-SNP pipeline using Treg-specific Hi-C, ATAC-seq, promoter and enhancer annotation datasets to identify interacting regulatory regions of the genome that are active within accessible chromatin in Tregs. Key to this process was the inclusion of the regulome of the master transcription regulator of Treg cells FOXP3, using human Treg specific ChIP information, given that FOXP3-binding is a major contributor to Treg function.

Using this pipeline, we identified 36 SNPs (from 1,228 T1D fine-mapped variants) with connectivity to new target genes via 3D chromatin interactions, 26 of which were interacting with enhancer regions. Crucially, we also identify another 119 interacting regions impacting 51 genes that may be involved in the disease. Connections between these new genes and variants have not been previously reported, demonstrating the power of cell type-specific interaction data in identifying novel disease risk in autoimmune disease. We further demonstrated the utility of the workflow by applying it to three other autoimmune datasets, identifying 16 more Treg-centric candidate SNPs and 35 interacting genes in different autoimmune disorders. Finally, we demonstrate the broad utility of the 3DFAACTS-SNP workflow for functional annotation of any genetic variation datasets by applying the filtering approach to ~2 million common (>10% allele frequency in populations) SNPs from the Genome Aggregation Database (gnomAD). In total, we found 7,900 SNPs and 3,245 candidate target genes, generating a list of potential sites for future T1D or autoimmune research.

As a data integration-based approach, one inevitable limitation of the 3DFAACT-SNP pipeline is its dependence on the data used to perform integration, particularly the quality and resolution of chromatin interactions identified from Hi-C data. In Chapter 3, instead of the standard analysis, which maps alignment read pairs to fixed-size genomic bins and using normalised contacts between bins as the chromatin interactions, we employed the post-filtered alignment read pairs from two Treg-specific Hi-C datasets as the basis of our chromatin interactions. Given the limited coverage of our initial Hi-C datasets, standard analyses would likely use large bin sizes (greater than 40 kb) due to the sparsity of the data, leading to an increase in false positive connections to regulatory elements. With more sequencing depth, such as billions of raw sequencing reads, we can generate contact maps with resolution at kilobases level (Rao *et al.*, 2014). In such cases, normalised contacts between fixed-size genomic bins with a small bin size such as 1 kb or 2 kb can more accurately indicate the real chromatin interactions, allowing a more accurate interpretation of the linkages between SNPs and regulatory regions, identifying more potentially causative SNPs for diseases. Therefore, in the chapter 4 and 5, we used published Hi-C datasets with high sequencing coverage to generate contact maps with 10 kb bin size, generating chromatin interactions with better validity.

The power of the 3DFAACTS-SNP pipeline is its ability to incorporate chromosome organisation in 3D as well as open chromatin annotation and detect functional chromatin interactions involving SNP-containing regulatory regions, leading to the discovery of *bona fide* target genes that have not previously been identified. While we initially used Treg-specific data and T1D-associated SNPs as a model to identify Treg-centric 3DFAACTS SNPs and their target genes, we have demonstrated that chromatin interactions from Hi-C dataset can be functionally mapped with multiple disease datasets as well as whole genome variant datasets such as variants from gnomAD, which presents a valuable resource in establishing cell-type specific interactomes. In our Treg-centric

3DFAACTS-SNP pipeline, we reasoned that Treg play an important role in the development and progression of autoimmune diseases so that Treg-specific epigenomics data must be used to model the potential causative regulation of autoimmune diseases. A similar case was recently applied to type 2 diabetes (T2D), where pancreatic islet tissue-centric epigenomics data were used (Greenwald *et al.*, 2018), demonstrating the power of using cell/tissue type-specific epigenomics data to interpret non-coding variants in complex diseases. The 3DFAACTS-SNP pipeline can be used as a standard post-GWAS analysis, to further prioritise variants with potential regulatory functionality to contribute to autoimmune diseases, hence reducing the cost of examining every variation identified by association studies for their functionality. Furthermore, this pipeline provides a useful mechanism to identify potential mechanisms by which non-coding variants regulate distal genes, allowing the discovery of novel diseases-associated target genes. These novel linkages and genes can eventually be used as potential targets for the development of novel diagnosis, prevention, and treatment plans of diseases.

‘Not all interactions are created equal’

From genetic, transcriptomic, and regulatory information, researchers are now able to go beyond a state of genetic linkage between genes and genetic variation, and truly link genetic changes to target genes on a functional level. Hi-C interaction contact maps of many cell lines and tissues have been investigated, and in Chapter 4 and 5 of this thesis, I access all publicly available Hi-C datasets to consolidate all interaction information into a single framework, implementing filtering and visualisation approaches that I describe in Chapter 2.

However, not all catalogued chromatin interactions, whether they be from one study or cell-type specific information, are biologically functional, and Hi-C datasets are prone to significant noise and proximity ligation can generate random physical connections that have no bearing on function. As we reviewed in Chapter 1, statistical model-based methods have developed to prioritise interactions that are more likely to be functional than others. For my thesis, I chose MaxHiC as my method to identify functional relevance, given it outperformed current existing models in identifying interactions with enrichment of regulatory elements such as promoters and enhancers (Alinejad-Rokny *et al.*, 2020). In Chapter 4, we therefore collected publicly available Hi-C datasets followed by analysing 173 datasets with the customized analysis pipeline, cataloguing a landscape of **Ma**xHiC-**d**etected **s**tatistically **si**gnificant **i**nteraction (MADSSI) profiles across 51 human cell lines and tissues. I also found that 62.3% of the MADSSI are uniquely found in only one cell line/tissue and enriched for cell/tissue-specific gene ontology (GO) terms, implying that the cell line/tissue-unique interactions are highly involved in cell/tissue-type specific gene regulation. Such unique interactions can be important differential marks for genomic regions may behave differently across cell and tissue types, providing an extra layer of information when associating cell/tissue-specific genetic variation, such as tissue-specific eQTLs, to the regulatory mechanism that is dysregulated in disease systems.

Additionally, by accumulating common (found in multiple cells and tissues) interactions, we defined regions that are found to have statistically significant interactions in more than half (at least 26) of all cell lines/tissues as interaction “hot zones”. We therefore identify 2,442 interaction hot zones, which were found to be significantly enriched for regulatory signals, such as active histone modification markers H3K27ac and H3K4me1, candidate cis-regulatory elements (cCREs) enhancer signatures. More interestingly, they are mostly enriched for insulator CTCF-binding sites and found to be located close to topologically-associated domains (TADs) boundaries, indicating the regulatory and structural functionality of the interaction hot zones. Compared to TADs, or sub-TADs, which are identified by statistical models from single dataset, the hot zones are identified based on the information across many cells and tissues, hot zones may provide additional structural marker information for a more accurate interpretation and investigation of the 3D genome architecture and structure-governed regulations that associated with diseases. Furthermore, other model organisms can potentially use the interaction hot zones as their 3D regulatory markers via alignment tools such as BLAST (Johnson *et al.*, 2008), facilitating the investigations of 3D gene regulation with limited data.

Subsequently, in Chapter 5, we annotated on average 75.35% cell/tissue type-specific MADSSI and comprehensively classified them into 66 interaction classes using epigenome annotations, revealing that MADSSI are mostly annotated to genes and active enhancers. Furthermore, we focused on interaction classes by reasoning expressed gene promoters and enhancers are functional regulatory markers, defining potentially regulatory functional MADSSI (PROF-MADSSI). Interestingly, on average 69.6% of them are enhancer-enhancer interactions, 49.6% are enhancer-CTCF binding site interactions and 21.5% are promoter-enhancer interactions, suggesting some interactions are identified as multiple types and overlaid by regulatory elements. More importantly, using the identified PROF-MADSSI, we used an example of PROF-MASSI contacting the *STAT6* gene promoter to demonstrate the power of using PROF-MADSSI to reveal regions enriched for regulatory elements. Therefore, in future studies investigating the gene regulation in different cells/tissues, the identified cell/tissue-specific PROF-MADSSI can be used to prioritise significant regions to be explored.

The GTEx project catalogued *cis* eQTLs of 54 human primary tissues (Lonsdale *et al.*, 2013), however it remains challenging to interpret the eQTLs because the mechanism by which eQTLs affect gene expression is unknown. In Chapter 5, we generated lists of cell/tissue-specific 3D regulatory regions, where regulatory annotations are contacted by the promoters of cell/tissue-specific expressed genes. We discovered an average of 26.73% these regions overlap with tissue-specific *cis* eQTLs. More importantly, we defined tissue-specific 3DeQTLs, which are eQTLs located within the 3D regulatory regions and contacting the promoter of their eQTL-target genes. We identified on average 5.9% of these 3D regulatory regions have 3DeQTLs interactions with the promoter of their eQTL-target genes, suggesting the unknown mechanism of how eQTLs affect gene expression is by affecting the 3D regulatory regions of that gene or altering the chromatin interactions between the target gene promoter and 3D regulatory regions.

Taken together, we demonstrated that using statistical model methods such as MaxHiC we can prioritise chromatin interactions that are potentially regulatory functional from publicly available Hi-C datasets, and successfully catalogue useful resources of cell and tissue-specific regulatory interactions and regions. Despite using unsupervised method such as kernel Principal Component Analysis (kPCA) to minimise the biases introduced by different samples when generating cell line/tissue-specific profiles, we cannot neglect systematic biases between studies, particularly the choice of Hi-C protocol and uneven sequencing depths of samples from different cells and tissues. For example, GM12878 and IMR90 were deeply sequenced in a number of studies (Dixon *et al.*, 2012; Rao *et al.*, 2014; Mifsud *et al.*, 2015), resulting in more statistically significant interactions being identified in these cell lines compared with others, obstructing fair comparisons between each cell and tissue. Another inevitable limitation is the matching efficiency between the annotations from public databases and interactions when we annotate interactions and categorise them into different interaction types. This is particularly limiting on the accuracy of profiling tissue-specific interaction types because compared with purified cells, different parts of a tissue may exhibit different 3D chromosome structure with different regulatory networks and gene expression patterns. Without such limitations, we can generate more accurate annotation profiles of the chromatin interactions, which will largely facilitate future studies of validating the regulations governed by 3D genome structure, and ultimately developing novel treatment or diagnosis approaches for diseases based on the validated regulations.

Future directions

In recent years, single cell sequencing methods such as single cell RNA-seq (scRNA-seq) (Tang *et al.*, 2009, 2019; Sasagawa *et al.*, 2013; Haque *et al.*, 2017) and ATAC-seq (scATAC-seq) (Lareau *et al.*, 2019; Satpathy *et al.*, 2019; Fang *et al.*, 2021) have driven the research on gene regulation mechanisms at the single cell level. More importantly, this has facilitated our understanding of cell type functional heterogeneity and the discovery of many sub-cell types with distinct gene expression profiles. Based on the same idea of incorporating single cell techniques, single cell Hi-C (scHiC) has been developed to profile chromatin interactions at a similar level (Nagano *et al.*, 2013; Stevens *et al.*, 2017; Tan *et al.*, 2018), implicating the necessity of investigating 3D chromatin interactions at single-cell level in the future in order to have a better understanding of the non-coding 3D regulations in different tissues and organs. However, compared with Hi-C, which constructs 3D chromosome structures using millions of cells, scHiC often generates data with high sparsity and noise, challenging the analysis and interpretation of the data (Zhou *et al.*, 2019; Kim *et al.*, 2020). Additionally, single-cell epigenomic data may not be available to conduct the same analyses that I present here in this thesis. Improved protocols and computational analysis approaches, such as sensitive unsupervised clustering algorithms and statistical models for sparse data, will therefore be needed specifically designed for scHiC datasets in order to accurately reveal causality between genetic risk and target genes.

While my work has shown links between regulatory mechanisms and target genes, significant work needs to be carried out to validate each specific 3D interaction. For example, we discovered novel linkages between T1D-associated SNPs-located in enhancer regions and promoters of the genes *CCR2*, *CCR3* and *CCR5* (Figure 3 in Chapter 3), suggesting that target genes altered functionality caused by these SNPs are dependent on 3D chromosome structural regulation. To validate this finding, enhancer region-knockout experiments could be developed in parallel with SNP editing experiments using CRISPR gene editing (Doudna and Charpentier, 2014), validating not only the connectivity between the enhancer and the gene expression, but also the impact of each SNP on expression of their target genes in *in vitro* models. Furthermore, approaches such as deep mutational scanning (DMS) (Fowler and Fields, 2014) and other multiplex assays for variant effect (MAVEs) (Kinney and McCandlish, 2019), could be used to quantify phenotypic effects for millions of genotypic variations in parallel, enabling the validation of the direct association between non-coding SNPs, 3D regulation and diseases. The validated non-coding SNPs regulating genes to contribute to diseases can then be used as a potential target for researchers and clinicians to develop novel diagnosis and treatment plans to provide alternative options for patients.

The identification of novel 3D interactions between non-coding variants to disrupted gene regulation in diseases systems offers an enormous promise to a “precision medicine” future, where precise diagnostics and treatment plans are developed at higher resolution and accounting for non-causal variations between individuals (Bainbridge *et al.*, 2011; Worthey *et al.*, 2011; Ashley, 2016; Ahmed, 2020; Morello *et al.*, 2020). For instance, Hi-C has revealed functional insights into cancer-specific risk loci (Jäger *et al.*, 2015; Du *et al.*, 2016; Hoskins *et al.*, 2016; Baxter *et al.*, 2018; Zhizhuo Zhang *et al.*, 2019), with such information being used to differentiate individual patients and allowing more accurate diagnosis models. Additionally, regulatory regions, particularly enhancers and super-enhancers have been demonstrated to play an important part in tumorigenesis (Donati, Lorenzini and Ciarrocchi, 2018; Gelato *et al.*, 2018; He, Long and Liu, 2019). In the future, the 3D regulatory regions including enhancers and super-enhancers identified in individual patients can serve as candidate targets for patient-specific enhancer therapies development.

References

Ahmed, Z. (2020) ‘Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis’, *Human genomics*, 14(1), p. 35.

Alinejad-Rokny, H. *et al.* (2020) ‘MaxHiC: robust estimation of chromatin interaction frequency in Hi-C and capture Hi-C experiments’, *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/2020.04.23.056226v1.abstract.

Ashley, E. A. (2016) ‘Towards precision medicine’, *Nature reviews. Genetics*, 17(9), pp. 507–522.

Atkinson, M. A., Eisenbarth, G. S. and Michels, A. W. (2014) ‘Type 1 diabetes’, *The Lancet*, 383(9911), pp. 69–82.

Ay, F., Bailey, T. L. and Noble, W. (2014) ‘Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts’, *Genome research*, 24(6), pp. 999–1011.

Bainbridge, M. N. *et al.* (2011) ‘Whole-genome sequencing for optimized patient management’, *Science translational medicine*, 3(87), p. 87re3.

Baxter, J. S. *et al.* (2018) ‘Capture Hi-C identifies putative target genes at 33 breast cancer risk loci’, *Nature communications*, 9(1), p. 1028.

Bernstein, B. E. *et al.* (2010) ‘The NIH Roadmap Epigenomics Mapping Consortium’, *Nature biotechnology*, 28(10), pp. 1045–1048.

Carty, M. *et al.* (2017) ‘An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data’, *Nature communications*, 8(1), p. ncomms15454.

Choi, S. W., Mak, T. S.-H. and O’Reilly, P. F. (2020) ‘Tutorial: a guide to performing polygenic risk score analyses’, *Nature protocols*, 15(9), pp. 2759–2772.

Davis, C. A. *et al.* (2018) ‘The Encyclopedia of DNA elements (ENCODE): data portal update’, *Nucleic acids research*, 46(D1), pp. D794–D801.

Dixon, J. R. *et al.* (2012) ‘Topological domains in mammalian genomes identified by analysis of chromatin interactions’, *Nature*, 485(7398), pp. 376–380.

Donati, B., Lorenzini, E. and Ciarrocchi, A. (2018) ‘BRD4 and Cancer: going beyond transcriptional regulation’, *Molecular cancer*, 17(1), p. 164.

Doudna, J. A. and Charpentier, E. (2014) ‘The new frontier of genome engineering with CRISPR-Cas9’, *Science*. Available at: https://science.sciencemag.org/content/346/6213/1258096.abstract?casa_token=y-Vudrweyd8AAAAA:fyFKGKFmZL0hNReyBnt2Nl-nHbCjmKhJrxMcawKC05NVwkCFNt3T1R6FQeM874VKXpC2BUtZOFif4g.

Du, M. *et al.* (2016) ‘Chromatin interactions and candidate genes at ten prostate cancer risk loci’, *Scientific reports*, 6, p. 23202.

ENCODE Project Consortium (2012) ‘An integrated encyclopedia of DNA elements in the human genome’, *Nature*, 489(7414), pp. 57–74.

Fang, R. *et al.* (2021) ‘Comprehensive analysis of single cell ATAC-seq data with SnapATAC’, *Nature communications*, 12(1), p. 1337.

Fernández, M. and Miranda-Saavedra, D. (2012) ‘Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines’, *Nucleic acids research*, 40(10), p. e77.

Fowler, D. M. and Fields, S. (2014) ‘Deep mutational scanning: a new style of protein science’, *Nature methods*, 11(8), pp. 801–807.

Freedman, M. L. *et al.* (2011) ‘Principles for the post-GWAS functional characterization of cancer risk loci’, *Nature genetics*, 43(6), pp. 513–518.

Gelato, K. A. *et al.* (2018) ‘Super-enhancers define a proliferative PGC-1α-expressing melanoma subgroup sensitive to BET inhibition’, *Oncogene*, 37(4), pp. 512–521.

Gorlov, I. P. *et al.* (2009) ‘GWAS meets microarray: are the results of genome-wide association studies and gene-expression profiling consistent? Prostate cancer as an example’, *PloS one*, 4(8), p. e6511.

Greenwald, W. W. *et al.* (2018) ‘Pancreatic islet chromatin accessibility and conformation defines distal enhancer networks of type 2 diabetes risk’, *BioRxiv*, p. 299388.

Hamdan, F. H. and Johnsen, S. A. (2019) ‘Perturbing Enhancer Activity in Cancer Therapy’, *Cancers*, 11(5). doi: 10.3390/cancers11050634.

Haque, A. *et al.* (2017) ‘A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications’, *Genome medicine*, 9(1), p. 75.

Hariprakash, J. M. and Ferrari, F. (2019) ‘Computational Biology Solutions to Identify Enhancers-target Gene Pairs’, *Computational and structural biotechnology journal*, 17, pp. 821–831.

He, Y., Long, W. and Liu, Q. (2019) ‘Targeting Super-Enhancers as a Therapeutic Strategy for Cancer Treatment’, *Frontiers in pharmacology*, 10, p. 361.

Hoskins, J. W. *et al.* (2016) ‘Functional characterization of a chr13q22.1 pancreatic cancer risk locus reveals long-range interaction and allele-specific effects on DIS3 expression’, *Human molecular genetics*, 25(21), pp. 4726–4738.

Huo, Y. *et al.* (2019) ‘Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk’, *Nature communications*, 10(1), p. 670.

International Multiple Sclerosis Genetics Consortium (IMSGC) *et al.* (2013) ‘Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis’, *Nature genetics*, 45(11), pp. 1353–1360.

Jäger, R. *et al.* (2015) ‘Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci’, *Nature communications*, 6, p. 6178.

Jin, F. *et al.* (2013) ‘A high-resolution map of the three-dimensional chromatin interactome in human cells’, *Nature*, 503(7475), pp. 290–294.

Johnson, M. *et al.* (2008) ‘NCBI BLAST: a better web interface’, *Nucleic acids research*, 36(Web Server issue), pp. W5–9.

Kerem, B. *et al.* (1989) ‘Identification of the cystic fibrosis gene: genetic analysis’, *Science*, 245(4922), pp. 1073–1080.

Kikuchi, M. *et al.* (2019) ‘Enhancer variants associated with Alzheimer’s disease affect gene expression via chromatin looping’, *BMC medical genomics*, 12(1), p. 128.

Kim, H.-J. *et al.* (2020) ‘Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data’, *PLoS computational biology*, 16(9), p. e1008173.

Kinney, J. B. and McCandlish, D. M. (2019) ‘Massively Parallel Assays and Quantitative Sequence–Function Relationships’, *Annual review of genomics and human genetics*. doi: 10.1146/annurev-genom-083118-014845.

de Lange, K. M. *et al.* (2017) ‘Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease’, *Nature genetics*, 49(2), pp. 256–261.

Lareau, C. A. *et al.* (2019) ‘Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility’, *Nature biotechnology*, 37(8), pp. 916–924.

Leinonen, R. *et al.* (2011) ‘The European Nucleotide Archive’, *Nucleic acids research*, 39(Database issue), pp. D28–31.

Lewis, C. M. and Vassos, E. (2020) ‘Polygenic risk scores: from research tools to clinical instruments’, *Genome medicine*, 12(1), p. 44.

Liang, B. *et al.* (2020) ‘GWAS in cancer: progress and challenges’, *Molecular genetics and genomics: MGG*, 295(3), pp. 537–561.

Lieberman-Aiden, E. *et al.* (2009) ‘Comprehensive mapping of long-range interactions reveals folding principles of the human genome’, *Science*, 326(5950), pp. 289–293.

Lizio, M. *et al.* (2015) ‘Gateways to the FANTOM5 promoter level mammalian expression atlas’, *Genome biology*, 16, p. 22.

Lizio, M. *et al.* (2019) ‘Update of the FANTOM web resource: expansion to provide additional transcriptome atlases’, *Nucleic acids research*, 47(D1), pp. D752–D758.

Lonsdale, J. *et al.* (2013) ‘The Genotype-Tissue Expression (GTEx) project’, *Nature genetics*, 45(6), pp. 580–585.

Mifsud, B. *et al.* (2015) ‘Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C’, *Nature genetics*, 47(6), pp. 598–606.

Mifsud, B. *et al.* (2017) ‘GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data’, *PloS one*, 12(4), p. e0174744.

Morello, G. *et al.* (2020) ‘From Multi-Omics Approaches to Precision Medicine in Amyotrophic Lateral Sclerosis’, *Frontiers in neuroscience*, 14, p. 577755.

Nagano, T. *et al.* (2013) ‘Single-cell Hi-C reveals cell-to-cell variability in chromosome structure’, *Nature*, 502(7469), pp. 59–64.

Pennacchio, L. A. *et al.* (2013) ‘Enhancers: five essential questions’, *Nature reviews. Genetics*, 14(4), pp. 288–295.

Rao, S. S. P. *et al.* (2014) ‘A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping’, *Cell*, 159(7), pp. 1665–1680.

Roadmap Epigenomics Consortium *et al.* (2015) ‘Integrative analysis of 111 reference human epigenomes’, *Nature*, 518(7539), pp. 317–330.

Royer-Pokora, B. *et al.* (1986) ‘Cloning the gene for an inherited human disorder—chronic granulomatous disease—on the basis of its chromosomal location’, *Nature*, 322(6074), pp. 32–38.

Sasagawa, Y. *et al.* (2013) ‘Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity’, *Genome biology*, 14(4), p. R31.

Satpathy, A. T. *et al.* (2019) ‘Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion’, *Nature biotechnology*, 37(8), pp. 925–936.

Stevens, T. J. *et al.* (2017) ‘3D structures of individual mammalian genomes studied by single-cell Hi-C.’, *Nature*, 544(7648), pp. 59–64.

Strathdee, C. A. *et al.* (1992) ‘Cloning of cDNAs for Fanconi’s anaemia by functional complementation’, *Nature*, 358(6385), p. 434.

Suna, O.-G. *et al.* (2015) ‘Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers’, *Nature genetics*, 47(4), pp. 381–386.

Tak, Y. G. and Farnham, P. J. (2015) ‘Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome’, *Epigenetics & chromatin*, 8, p. 57.

Tang, F. *et al.* (2009) ‘mRNA-Seq whole-transcriptome analysis of a single cell’, *Nature methods*, 6(5), pp. 377–382.

Tang, X. *et al.* (2019) ‘The single-cell sequencing: new developments and medical applications’, *Cell & bioscience*, 9, p. 53.

Tan, L. *et al.* (2018) ‘Three-dimensional genome structures of single diploid human cells’, *Science*, 361(6405), pp. 924–928.

Worthey, E. A. *et al.* (2011) ‘Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease’, *Genetics in medicine: official journal of the American College of Medical Genetics*, 13(3), pp. 255–262.

Zhang, Z. *et al.* (2019) ‘An AR-ERG transcriptional signature defined by long-range chromatin interactomes in prostate cancer cells’, *Genome research*, 29(2), pp. 223–235.

Zhang, Z. *et al.* (2019) ‘Transcriptional landscape and clinical utility of enhancer RNAs for eRNA-targeted therapy in cancer’, *Nature communications*, 10(1), p. 4562.

Zhou, J. *et al.* (2019) ‘Robust single-cell Hi-C clustering by convolution- and random-walk–based imputation’, *Proceedings of the National Academy of Sciences of the United States of America*, 116(28), pp. 14011–14018.