UNIVERSITY OF CALIFORNIA SAN DIEGO

Exploring the Regulatory Genome and Functional Genetic Variation

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

William Walter Young Greenwald

Committee in charge:

Professor Kelly Frazer, Chair
Professor Vineet Bafna
Professor Melissa Gymrek
Professor Olivier Harismendy
Professor Graham McVicker

2019

The Dissertation of William Walter Young Greenwald is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

DEDICATION

To my parents, Kay and Lee, and my wife, Amanda.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Kelly Frazer and Dr. Erin Smith as mentors throughout my doctoral process. I would not be the scientist that I am today without their guidance.

Chapter 1, in full, is a reprint of the material as it appears in BMC Bioinformatics, 2017, William W. Greenwald, He Li, Erin N. Smith, Paola Benaglio, Naoki Nariai, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Nature Communications, 2019, William W. Greenwald, He Li, Paola Benaglio, David Jakubosky, Hiroko Matsui, Anthony Schmitt, Siddarth Selvaraj, Matteo D'Antonio, Agnieszka D'Antonio-Chronowska, Erin N. Smith, Kelly A. Frazer. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 3, in full, has been submitted for publication of the material as it may appear in Nature Genetics 2019, William W. Young Greenwald, Agnieszka D'Antonio-Chronowska, Paola Benaglio, Hiroko Matsui, Erin N. Smith, Matteo D'Antonio, Kelly A. Frazer. The dissertation author was one of the primary investigators and authors of this paper.

VITA

2011-2015 Bachelor of Science, Computational and Systems Biology, University of California Los Angeles

2015-2019 Doctor of Philosophy, Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

First Author

Chromatin co-accessibility is highly structured, spans entire chromosomes, and mediates long range regulatory genetic effects. *In Review: Nature Genetics.* 2019

Pancreatic islet chromatin accessibility and conformation defines distal enhancer networks of type 2 diabetes risk. *Accepted: Nature Communications.* 2019

Subtle Changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nature Communications.* 2019

Utilization of defined microbial communities enables effective evaluation of meta-genomic assemblies. *BMC Genomics.* 2017

Pgltools: a genomic arithmetic tool suite for manipulation of Hi-C peak and other chromatin interaction data. *BMC Bioinformatics*. 2017

Other Authorship

Allele-specific NKX2-5 binding underlies multiple genetic associations with human EKG traits. *In Review: Nature Genetics*

Human iPSC-derived retinal pigment epithelium: a model system for identifying and functionally characterizing causal variants at AMD risk loci. *In Review: Stem Cell Reports*

Identification of common and rare genetic variation associated with plasma protein levels using whole-exome sequencing and mass spectrometry. *Circulation: Genomic and Precision Medicine*. 2018

Insights into the Mutational Burden of Human Induced Pluripotent Stem Cells from an Integrative Multi-Omics Approach. *Cell Reports*. 2018

Updated and standardized genome-scale reconstruction of Mycobacterium Tuberculosis H37Rv, iEK1011, simulates flux states indicative of physiological conditions. *BMC Systems Biology*. 2018

Efficient prioritization of multiple causal eQTL variants via sparse polygenic modeling. *Genetics.* 2017

iPSCORE: A resource of 222 iPSC lines enabling functional characterization of genetic variation across a variety of cell types. *Stem Cell Reports.* 2017

Decreased STARD10 expression is associated with defective insulin secretion in humans and mice. *American Journal of Human Genetics.* 2017

ABSTRACT OF THE DISSERTATION

Exploring the Regulatory Genome and Its Contained Genetic Variation

by

William Walter Young Greenwald

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2019

Professor Kelly Frazer, Chair

A substantial fraction of SNPs associated with human traits and diseases through genome-wide association studies (GWAS) are likely regulatory variants as they tend to

be located within enhancers and associated with differential gene expression. Thus, as a key step in implementing personalized medicine, it is important to identify regulatory variants in the human genome, and characterize their underlying molecular mechanisms. However, identifying and elucidating the functions of regulatory variants is currently challenging as these variants show similar associations with many other neutral variants due to linkage disequilibrium, can be quite far from the gene(s) they regulate, and often have cell type-specific effects. In order to overcome these challenges and interrogate the function of these regulatory variants, it could be possible to examine and integrate epigenetic information in a cell type dependent manner. Here, I present three studies which focus on the functionality of the epigenome – specifically chromatin looping and co-accessibility – in the context of gene regulation, genetics, and disease. I present a tool for computationally working with chromatin loop data, and utilize this tool to show that genetic variation is not associated with large changes in chromatin looping, but rather small modulation in contact propensity which are associated with large changes in gene expression. I then examine chromatin co-accessibility, and show that genetic variants may be able to mediate long range effects on genes and accessible sites hundreds of megabases away – across entire chromosomes – which are associated with cell type relevant genes and diseases.

CHAPTER 1 PGLTOOLS: A GENOMIC ARITHMETIC TOOL SUITE FOR MANIPULATION OF HI-C PEAK AND OTHER CHROMATIN INTERACTION DATA

Abstract

Background: Genomic interaction studies use next-generation sequencing (NGS) to examine the interactions between two loci on the genome, with subsequent bioinformatics analyses typically including annotation, intersection, and merging of data from multiple experiments. While many file types and analysis tools exist for storing and manipulating single locus NGS data, there is currently no file standard or analysis tool suite for manipulating and storing paired-genomic-loci: the data type resulting from "genomic interaction" studies. As genomic interaction sequencing data are becoming prevalent, a standard file format and tools for working with these data conveniently and efficiently is needed.

Results: This article details a file standard and novel software tool suite for working with paired-genomic-loci data. We present the paired-genomic-loci (PGL) file standard for genomic-interactions data, and the accompanying analysis tool suite "pgltools": a cross platform, pypy compatible python package available both as an easy-

to-use UNIX package, and as a python module, for integration into pipelines of paired-genomic-loci analyses.

Conclusions: Pgltools is a freely available, open source tool suite for manipulating paired-genomic-loci data. Source code, an in-depth manual, and a tutorial are available publicly at www.github.com/billgreenwald/pgltools, and a python module of the operations can be installed from PyPI via the PyGLtools module.

Background

Numerous experimental methodologies have been developed in the past decade to study 3D configurations of the human genome, including Hi-C and ChIA-PET[1,2]. These "genomic interaction" data have provided key insights into the regulation of gene expression, and suggest that chromatin interactions are driven by discrete, yet spatially-associated, epigenetic features[3,4]. File standards and tool suites have become essential to conduct efficient bioinformatics analyses; for example, single locus information can be encoded in the BED file format and manipulated using bedtools, enabling a wide variety of bioinformatics inquiries[5]. However, it is currently challenging to fully interpret the biological impact of genomic interactions as tools do not yet exist to quickly and iteratively interrogate the extent to which both regions of paired loci are conserved across genomic datasets from diverse cell-types and contexts. While paired-genomic-loci data generated from these methodologies are widely available, the bioinformatics field has not yet developed either a file standard or analysis tools for their efficient manipulation.

There are currently several file formats for paired-genomic-loci, however, none of these file formats were designed to enable efficient annotation and data manipulation. Existing file formats include those that encode read count information such as the matrix and the triplet sparse matrix formats[6], and others that encode the locations of paired segments and specialized metadata for particular pipelines, such as the HiFive ChromatinInteraction format[7]. Although the matrix and triplet sparse matrix formats effectively communicate coverage depth across bins of the genome, they are restricted to fixed locus bin sizes, are not human-readable, and are cumbersome for genomic arithmetic. Additionally, while the ChromatinInteraction format, and the similarly structured bedtools bedpe format[5], may appear to be suitable storage formats for integration into a genomic arithmetic pipeline, as the two loci can be written in any order within the file, programmatic manipulation is unnecessarily complicated. Finally, the triplet sparse matrix and ChromatinInteraction formats are both specialized for the specific programs for which they were designed. Thus, to facilitate genomic interaction data manipulation, allow for variable locus bin sizes within a single data set, and allow for flexible metadata important to paired-genomic-loci, a new file format standard is needed.

Numerous analysis tools exist to process, normalize, or call peaks from raw reads of paired-genomic-loci data[3,6-9], yet there is no software that performs efficient manipulation and genomic arithmetic, analogous to bedtools, for single locus data, hindering the process of annotating and comparing chromatin interactions. For example, bedtools does not provide operations for bedpe that analyze both loci simultaneously, and

there are no tools for genomic arithmetic within HiFive. Furthermore, a tool for converting to the ChromatinInteraction format, or for converting from the triplet sparse matrix format to visualization formats, does not currently exist. An analysis tool suite that performs efficient manipulation and genomic arithmetic of paired-genomic-loci data would allow for more complete analyses of these datasets, and thus the potential to gain deeper biological insights about the 3D conformation of the human genome.

Here we describe a new file standard for paired-genomic-loci data, the PGL format, and an analysis tool suite, pgltools, for genomic interaction data storage and manipulation. The PGL format supports genomic interaction data, allows for appropriate metadata, and enables efficient data manipulation. Pgltools performs genomic arithmetic on PGL files such as comparing, merging, and intersecting two sets of paired-genomic-loci, as well as integrates BED files with PGL files. Finally, we provide functions to convert other genomic interaction file formats to PGL files, and convert PGL files to multiple different visualization formats. This analysis tool suite will allow for iterative bioinformatics analyses and visualization of genomic interaction data, facilitating discovery and collaboration within the genomic interaction field.

Implementation

Our goal was to create a file standard that can summarize the output from mapping and peak calling algorithms for chromatin interaction data derived from experiments, such as Hi-C or ChIA-PET, that is easily interpretable, shareable, and can be combined with current genomic annotation formats, such as the BED format. We first

established a paired-genomic-loci file standard—the “PGL” file type— which represents each paired-genomic-loci as a single PGL entry in a human readable text file, with space in each entry for annotations, and then implemented an analysis tool suite for working with these files. Within “genomic interaction” data, the interactions between two loci (locus A and locus B) are captured—this “paired” information is preserved through the PGL file standard. PGL files require six columns in the following order: locus A chromosome, locus A start position, locus A end position, locus B chromosome, locus B start position, and locus B end position. Beyond the six columns, any user-defined annotations, such as interaction p-value or locus chromatin state, can be written. These annotations can be manipulated and utilized by the operations in PGLtools to gain insight into the relationship between multiple paired-genomic-loci. As annotations are unique to a file, headers can be given in files by preceding a line with “#.” Furthermore, PGL files are required to have each PGL entry written such that locus A comes before locus B based on chromosome number alphabetically (ex. chr1, chr10, chr15, chr22, chr7, chrX, chrY) and chromosome position numerically. This relationship, when combined with file sorting on each column sequentially, gives pgltools the ability to quickly merge and intersect PGL entries from PGL files. Operations for sorting PGL files, converting files to PGL files, and formatting PGL files for visualization with established programs, are also included in pgltools.

Most pgltools operations utilize the same core function to test for overlapping paired-genomic-loci within or between file(s). For single locus entries, such as those in sorted BED files, overlapping entries must be sequential: if entries 1 and 3 overlap, entry 2 must overlap both entries 1 and 3 (Figure 1.1A). This property allows bedtools to limit

of the number of features that must be compared for overlap, thus expediting analyses[5]. However, in sorted PGL files, while locus A from multiple sequential entries can overlap, locus B may not overlap (Figure 1.1B). The pgltools *overlap* function allows for this and quickly and efficiently finds consecutive and non-consecutive entries where both locus A and locus B are overlapping. It begins by comparing the first PGLs in both files, recording if an overlap occurred in both loci, and then advances to the next PGL in File 2. These comparisons continue until the PGL from File 2 does not overlap locus A from the PGL in File 1, at which point the algorithm begins comparing the next PGL from File 1 to the first possible overlapping PGL from File 2. This repeats until the ends of both files are reached. An in-depth flow chart of the overlap operation's control flow, as well as how the first possible overlapping PGL from File 2 is determined, is shown in Figure 1C. Pgltools is implemented in Python 2.7, and all operations have been tested with the pypy python compiler. As such, the UNIX package version of pgltools can be run either with CPython or pypy; the included UNIX wrapper will run pgltools through pypy if installed, or CPython if pypy is not installed. Utilizing pypy reduces memory consumption by approximately 25%, and decreases run times 5-7 fold. The pgltools suite can read from UNIX standard in, useful for stringing multiple pgltools commands together without needing to save the intermediate files, and writes to UNIX standard out, allowing it to be utilized in complex pipelines to speed up analysis of genomic interaction data. Pgltools is also available as a python module, PyGLtools, for use within pythonic pipelines, and can be installed from PyPI. As pgltools is written in Python 2.7, it is easily portable to any platform and poised for collaboration with the community.

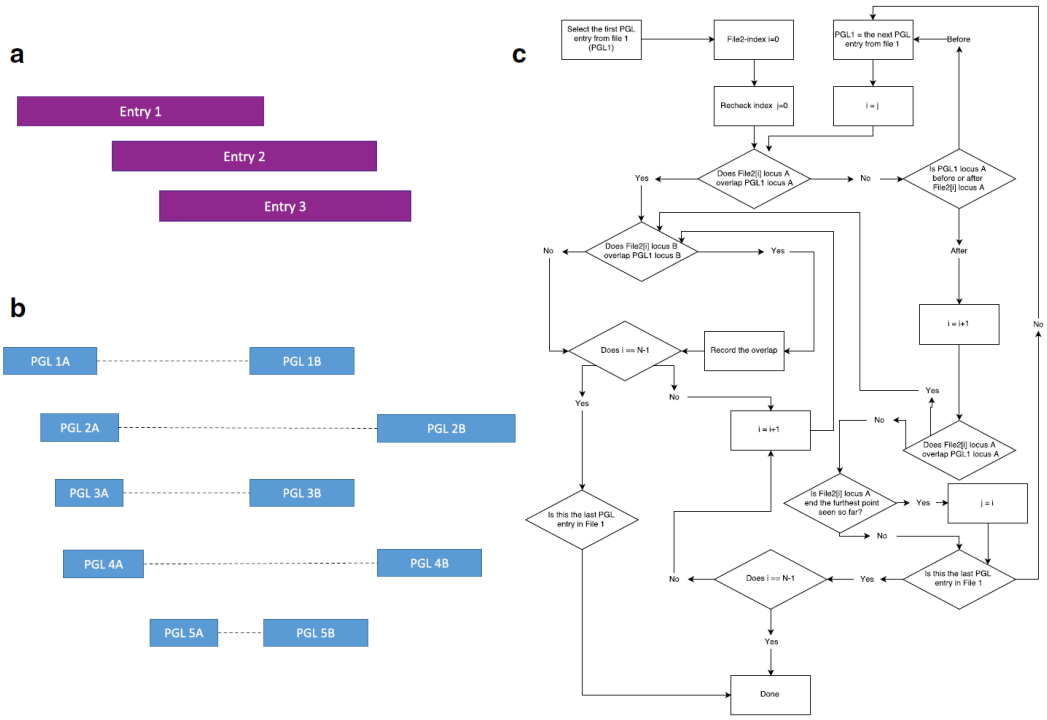

Figure 1.1 Pgltools Implementation
(a) An example of sorted, single locus bed file entries from a file sorted by start position. As entry1 overlaps entry 3, entry 2 must also overlap entry 3. (b) A pictorial representation of PGL entries in a sorted PGL file where non-sequential PGL entries overlap. Loci are shown as blocks, with *dashed lines* connecting the paired-loci comprising a single entry. Both loci A and B in PGL entries 1 and 3 overlap, and both loci in PGl entries 2 and 4 overlap. (c) A flowchart of the overlap function shared between many operations in pgltools. File 2 has N-1 entries. File 2 is iterated by the File2-index i. File2[i] is a PGL entry for any $0 \leq I < N$. Throughout the algorithm, PGL entries from File 2 must be checked multiple times. Therefore, to reduce the number of comparisons performed by pgltools, the Recheck Index is used to store the index at which the previous overlap iteration began. When the ends of both files are reached, the algorithm ends.

Results and Discussion

Table 1.1 includes a full list of pgltools operations and their default behavior. Visualizations of these operations are provided in Figure 1.2. The pgltools *intersect* operation can be used to identify the overlapping, union, or unique PGL entries between two PGL files, while preserving or combining annotations during these analyses; for example, the number of overlapping bases at each locus from each PGL entry from two PGL files can be determined. The pgltools *merge* operation can be utilized to merge

overlapping PGL entries, or PGL entries within a specified distance within a single PGL file. Summary statistics, such as the number of merged entries, can be obtained through command line arguments to the *merge* operation. To determine differential PGL entries between two PGL files, the *subtract* operation has been included to remove the parts of PGL entries present in one PGL file from those present in another. Once a set of PGL entries has been determined, it is common to filter these entries to a desired genetic region—the *window* operation can be used to filter based on either or both end(s) of the PGL entries in a PGL file. To interrogate questions regarding differential coverage depth of genomic interactions, such as genetic association with interaction intensity, we provide the *samTopgl* operation, which when utilized with the *coverage* operation, will find the number of reads from a sam file that overlap each PGL entry in a PGL file (though the operation is generalizable for any two PGL files). The *closest* operation is provided for finding the closest PGL entries between two PGL files. The *expand* operation can expand both loci by a given value. In addition, as single locus genomic metadata is often analyzed together with interaction data, such as presence of a coding region, epigenetic annotation, or motif locations, we provide *intersect1D, closest1D,* and *subtract1D* operations for analysis on traditional BED files and PGL files. Finally, we include helper operations both for converting files to the PGL format, including *formatbedpe* to convert a bedpe file and *formatTripSparse* to convert triple sparse matrix files, and for converting from the PGL format to packages for visualization or further analysis, such as the *conveRt* operation to convert to a file readable by the GenomicInteractions R package[10], *browser* for visualizing with the UCSC Genome Browser[11], *juiceBox* for visualizing with

JuiceBox[3,12], and *condense* and *findLoops* to create a BED file of either the anchors or interior regions of each PGL.

Table 1.1 Summary of operations provided in pgltools

Method	Description
intersect	Find overlapping paired-genomic-loci from two PGL files
merge	Merge nearby paired-genomic-loci within a single file and produce a column containing summary statistics requested through passed parameters (-c and -o)
subtract	Find part of paired-genomic-loci from a PGL file that do not overlap another PGL file
window	Filter a PGL file to a particular genomic region
samToPgl	Converts a SAM file to a PGL file
coverage	Find the coverage of a PGL file on another PGL file; usually used to find the coverage of reads from a PGL file derived from a SAM file on a set of PGLs. The paired-genomic-loci from file 2 only need to overlap the paired-genomic-loci from file 1.
closest	Find the closest paired-genomic-loci from a PGL file to each paired-genomic-loci in another PGL file
expand	Expand both loci by a given size
closest1D	Find the paired-genomic-loci that overlap regions from a bed file
subtract1D	Find the parts of paired-genomic-loci that do not overlap regions from a bed file
sort	Sorts a PGL file for use with other PGLtools operations
formatbedpe	Convert a bedpe-like file to a PGL file
formatTripSparse	Convert a triplet sparse matrix file set to a PGL file
conveRt	Formats the PGL file for use with the GenomicInteractions R package
browser	Format a PGL file to be viewed in the UCSC Genome Browser
juicebox	Format a PGL file to be viewed in juicebox
condense	Convert a PGL file to a BED file with two entries for each PGL entry
findLoops	Convert a PGL file to a BED file with an entry containing the region from the start of anchor A to the stop of anchor B for intra-chromosomal PGLs, and an entry for each anchor for inter-chromosomal PGLs.

By combining the operations within pgltools, one can quickly and easily interrogate biological functionality in the context of chromatin interaction data. For example, by combining the *intersect1D* and *merge* operations, it is possible to determine the different chromatin annotations for each locus of each PGL entry (which could then be further filtered to determine 3D interactions between chromatin states of interest, e.g. promoter-enhancer). Additionally, pgltools can be used to find overlaps between chromatin interactions and other types of paired data. For example, one could create a PGL file from a list of expression Quantitative Trait Loci (eQTLs) and their corresponding target genes (eGenes), and utilize the *intersect* operation to determine if any pairs of eQTL and eGenes fall within a chromatin interaction. Example pipelines for these scenarios can be found on the pgltools github.

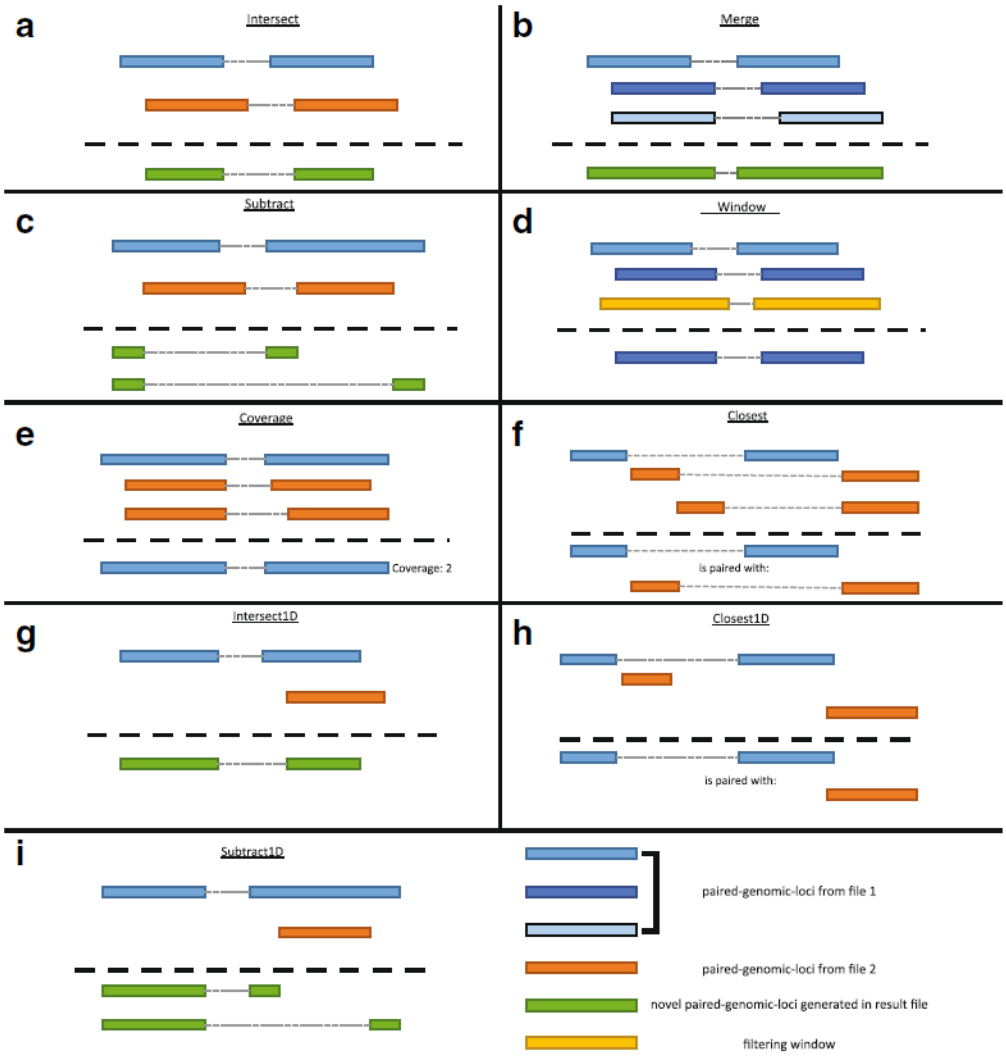

Figure 1.2 The operations of pgltools.
PGL entries from file one are shown in various shades of blue, PGL entries from file two are shown in orange, and windows are shown in yellow (see legend at bottom right). All resulting outputs are shown below dashed lines, with novel entries shown in green and original entries shown in their original color. (a) The intersect operation finds overlapping paired loci between two PGL files and returns the overlapping regions. (b) The merge operation combines overlapping paired loci within a single PGL file. (c) The subtract operation returns the PGL entries from file one with the PGL entries from file two removed. (d) The window operation returns the PGL entries that fall completely within a specified genomic region. (e) The coverage operation returns the number of PGL entries from file two that overlap each PGL entry in file one. (f) The closest operation returns the closest PGL entry from file two for each PGL entry in file one. (g) The intersect1D operation returns PGL entries from file one that overlap regions in a bed file. (h) The closest1D operation returns the closest region from a bed file for each PGL entry in file 1. (i) The subtract1D operation returns the PGL entry from file one with the regions from a bed file removed.

Conclusions

Pgltools is an open source software analysis tool suite for interacting with the PGL file standard for paired-genomic-loci. Pgltools can read from and writes to UNIX standard in and standard out, and can be run quickly in both CPython and pypy. A python module version, PyGLtools, is available for use within pythonic pipelines. The cross-platform nature of python poises pgltools for community contribution, and makes it easy to install and utilize.

Chapter 1, in full, is a reprint of the material as it appears in BMC Bioinformatics, 2017, William W. Greenwald, He Li, Erin N. Smith, Paola Benaglio, Naoki Nariai, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

CHAPTER 2 SUBTLE CHANGES IN CHROMATIN LOOP CONTACT PROPENSITY ARE ASSOCIATED WITH DIFFERENTIAL GENE REGULATION AND EXPRESSION

Abstract

While genetic variation at chromatin loops is relevant for human disease, the relationships between contact propensity (the probability that loci at loops physically interact), genetics, and gene regulation are unclear. We quantitatively interrogate these relationships by comparing Hi-C and molecular phenotype data across cell types and haplotypes. While chromatin loops consistently form across different cell types, they have subtle quantitative differences in contact frequency that are associated with larger changes in gene expression and H3K27ac. For the vast majority of loci with quantitative differences in contact frequency across haplotypes, the changes in magnitude are smaller than those across cell types; however, the proportional relationship between contact propensity, gene expression, and H3K27ac are consistent. These findings suggest that subtle changes in contact propensity have a biologically meaningful role in gene regulation and could be a mechanism by which regulatory genetic variants in loop anchors mediate effects on expression.

Introduction

Chromatin loops colocalize regulatory elements with their targets[2,3,13-25] by bringing genomic regions that are distant from one another in primary structure close together in 3D space[26]. These colocalized regions, also known as loop anchors, are

preferentially enriched for disease associated distal regulatory variation and expression quantitative trait loci (eQTLs)[27-32]. While it has been shown that the physical 3D distance between looped loci can vary[26,33-35], previous studies examining cell type and haplotype differences in looping have considered loops to be either present or absent, rather than a quantitative phenotype. Thus, the extent to which quantitative differences between chromatin loops exist, and whether they are associated with differences in gene expression and regulation, has yet to be explored.

Bulk chromatin conformation assays (e.g. 3C, 4C, and Hi-C) were designed to measure physical contact frequency between two pieces of colocalized (ie looped) DNA in a pool of cells. While a recent single cell Hi-C study found that contacts occur within single cells at loops called from bulk data, there was variability in the contact profiles of looped loci between cells[36]. Together, this suggests that the contact frequency measured in a pool of cells reflects the proportion of cells in which a contact is occurring, or the probability for the contact to occur (contact propensity) across all cells in the sample. Investigating contact frequency as measured by Hi-C, in combination with molecular phenotypes, may reveal if contact propensity between looped loci varies across cell types and haplotypes, and if this variation is associated with differential regulation of gene expression.

If contact propensity between looped loci does in fact play a role in gene regulation, a genetic variant that affects contact propensity would likely have a downstream effect on gene expression. Therefore, the association between contact

propensity and gene expression would exist not only across cell types, but also across haplotypes. Recent studies examining whether chromatin loops vary across haplotypes, and the functional consequences of this variation, have come to conflicting conclusions. Rao et al.[2] created and phased the GM12878 Hi-C map (which is the highest resolution map currently available) to study differences in looping across haplotypes, and did not observe differences between the paternal and maternal haplotypes outside of imprinted regions. Other more recent studies employing CTCF ChIA-PET[16] and H3K27ac Hi-ChIP[37] have reported that allelic imbalance in chromatin looping occurs throughout the genome. These contradictory results are likely due to the experimental design and types of effects examined in these studies. Rao et al.[3] used Hi-C data to look for large differences across haplotypes, and thus may have missed smaller effects. The studies using ChIA-PET and Hi-ChIP sought to identify allelic imbalance of all sizes, but employed experimental approaches that may be biased as they simultaneously measure either CTCF binding or regulatory region activity and chromatin looping, thereby conflating the allelic bias of the two phenotypes. A genome-wide quantitative analysis into allele specific chromatin looping using phased Hi-C would enable the unbiased estimation of the magnitude at which contact propensity varies across haplotypes at all types of chromatin loops (rather than only those at promoters and/or enhancers). Additionally, integrating this data with phased gene expression and H3K27ac data could provide evidence that contact propensity plays a role in in long-range gene expression regulation, and provide insight into how regulatory genetic variants may influence chromatin structure.

In this study, we generate a resource of phased, high resolution Hi-C chromatin maps from induced pluripotent stem cells (iPSCs) and iPSC-derived cardiomyocytes (iPSC-CMs) from seven individuals in a three-generation family for whom we have 50X whole genome sequence (WGS), and phase gene expression (RNA-seq) and enhancer activity (H3K27ac ChIP-seq) data generated from the same iPSCs and iPSC-CMs. We identify chromatin loops, quantitatively characterize cell type associated looping, and find that that while loops tend to be present in both cell types, some loops exhibit significantly increased contact propensity within one cell type. We show that these quantitatively-identified cell type associated loops (CTALs) recapitulate known biology discovered through previous qualitative comparisons of cell type specific loops, including being enriched for differentially expressed genes and regulatory regions, becoming more specialized throughout differentiation, and connecting distal eQTLs to their target gene. Additionally, our quantitative analyses reveal that small magnitude changes in contact propensity are proportionally associated with large changes in molecular phenotypes: an association that could not be identified by qualitative comparisons. We next examine allelic differences in contact propensity by phasing our Hi-C data, and find that haplotype associated chromatin loops (HTALs) are highly enriched for imprinted regions or for being associated with copy number variation, but not for eQTLs, suggesting that regulatory genetic variants do not exert large effects on chromatin contact propensity. Finally, we examine the association between differential contact propensity and differential gene expression and H3K27ac over a range of magnitudes across both cell types and haplotypes by quantitatively associating the phenotypes in aggregate across the genome. These analyses reveal a genome-wide proportional relationship between

differential contact propensity and differential expression and H3K27ac that is consistent across cell types and haplotypes. Our study therefore suggests that the cellular context of a chromatin loop (ie cell type, genetics, etc.) affects the propensity for an interaction at a loop to occur, and that these small changes to contact propensity are associated with large functional effects. This model suggests that regulatory genetic variation could mediate its effects on gene expression through subtle modification of contact propensity at chromatin loops.

Results

Sample and data collection

Molecular data was obtained from iPSCs and their derived cardiomyocytes (iPSC-CMs) from seven individuals in a three-generation family from iPSCORE (the iPSC collection for Omics REsearch)[38] (Figure 2.1A). Fibroblasts from these seven individuals were reprogrammed using non-integrative Sendai virus vectors[39], from which eleven iPSC lines were generated and subsequently differentiated into thirteen iPSC-CM samples using a monolayer-based protocol[40]. From the eleven iPSC and thirteen iPSC-CM samples, we generated chromatin interaction data via *in situ* Hi-C[3]. Additionally, from these and other iPSC and iPSC-CM samples from the same seven individuals, we integrated functional genomic data that was generated as part of a concurrent manuscript[41] (RNA-seq for gene expression, H3K27ac ChIP-seq for enhancer activity, and ATAC-seq for chromatin accessibility; Figure 1B; see methods) which also describes the differentiation efficiency and quality of all iPSC and iPSC-CM lines used in this study. Finally, we obtained single-nucleotide variants (SNVs) and somatic and inherited

copy-number variants (CNVs) for the seven individuals from ~50X WGS and genotype arrays from previously published work[38,42].

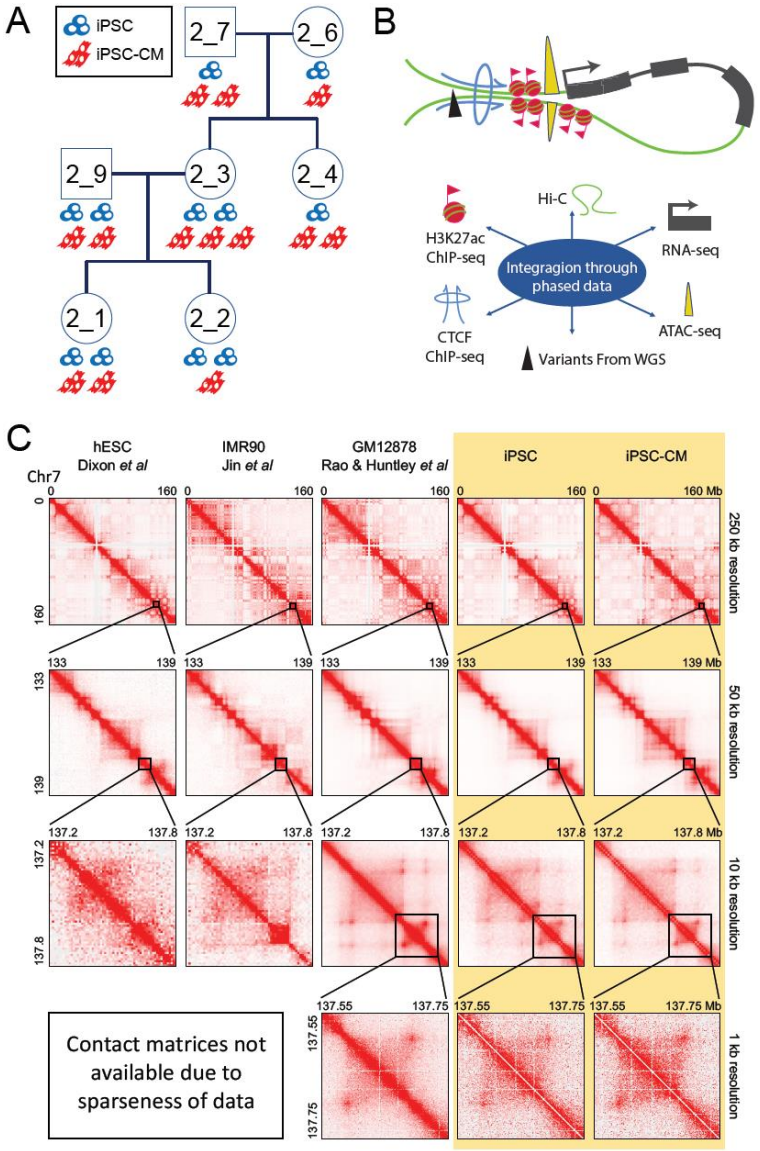

Figure 2.1 Study design, data, and chromatin contact maps
(A) Pedigree of the seven individuals used in this study. Cell icons below each subject indicate the number of iPSC lines and iPSC-CM samples used in the Hi-C experiments. iPSC lines are shown in blue, iPSC-CM samples are shown in red. (B) Schematic showing the data types used in this study depicting how they colocalize at loop anchors. (C) Hi-C contact maps from previous Hi-C studies (three left columns), and this study (two right columns, highlighted in yellow), displaying depth of map on Chromosome 7 (arbitrarily chosen for example). For Dixon *et al* and Jin *et al*, the data is too sparse to zoom to 1kb resolution.

Identification of chromatin loops in iPSCs and iPSC-CMs

We characterized the 3D chromatin structure of iPSCs and iPSC-CMs by identifying chromatin loops in each cell type genome-wide. From the *in situ* Hi-C data, we obtained 1.74 billion long-range (≥20kb) intra-chromosomal contacts after aligning and filtering ~6 billion Hi-C read pairs across all twenty-four Hi-C samples. We performed hierarchical clustering of the contact frequencies by cell type across individuals and observed high correlations within each cell type both by Pearson correlation, and by correcting for Hi-C biases via HiCRep[43]. To identify a set of reference loops for downstream quantitative analyses, we combined the Hi-C data within each cell type to obtain a comprehensive set of loops from high-depth data. We pooled the data across samples for each cell type, resulting in reference chromatin maps with the highest resolution (~2kb matrix resolution, defined as the resolution at which 80% of loci have 1000 or more contacts with any other locus[3]) in iPSCs and iPSC-CMs (or any other iPSC derived cell type) to date, and were comparable in resolution to the Hi-C map in GM12878[3] (Figure 2.1C). As loop calling algorithms often identify distinct loops, and are dependent on the resolution parameters specified for their analysis[44], we called chromatin loops from these maps utilizing two algorithms (HICCUPS and Fit-Hi-C) at multiple resolutions, identifying 17,567 loops in iPSCs (iPSC called loops), and 19,003 iPSC-CM loops (iPSC-CM called loops). We examined the overlap of the loop calls between cell types (Figure 2.2A) and found that 37.1% of the total 26,679 loops were called in both cell types (Figure 2.2B). These findings were consistent with previous studies investigating differential presence of loops between cell types[3,45]. To examine whether

these loops were predominantly demarcating TADs, or were separate from TAD structure, we also called TADs in both cell types and examined the number of loops that had both anchors within 25kb of TAD boundaries. We found only 2.9% of iPSC loops, and 5.1% of iPSC-CM loops, to have both anchors at TAD boundaries, indicating that these loops were primarily not demarcating TADs. These iPSC and iPSC-CM called loop sets provide a resource for the analysis of long-range gene regulation across the genome.

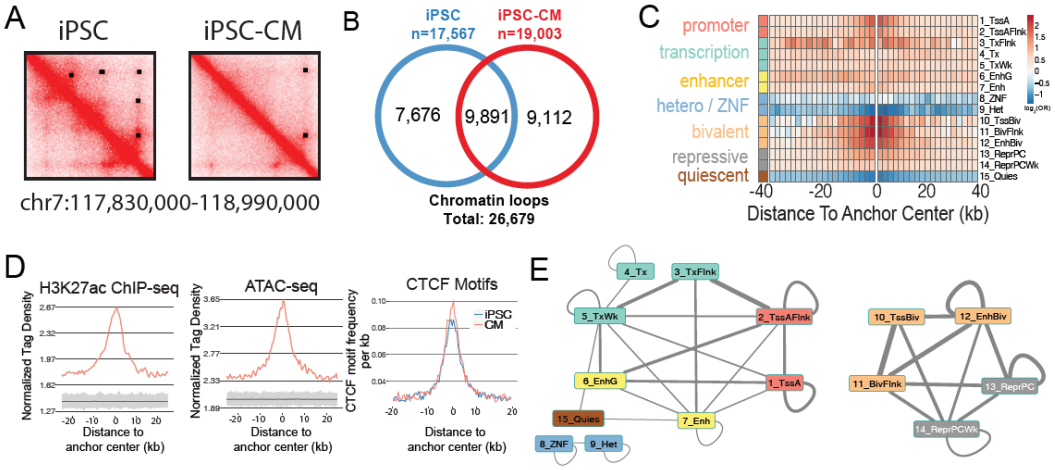

Figure 2.2 iPSC and iPSC-CM called loops
(A) Example contact maps from iPSCs (left) and iPSC-CMs (right) showing differences in looping identified by callers across cell types, with loop calls shown on the top right half of maps as black rectangles. Two loops appear present and are called in iPSCs, and 4 loops appear present and are called in iPSC-CMs. (B) Venn diagrams showing the number of chromatin loops unique and common to both cell types. (C) Heatmap showing enrichment of regulatory regions near iPSC-CM called loop anchor centers. The 15 ROADMAP chromatin states of fetal heart tissue (E083) were used, and the $\log_2$ odds ratio of enrichment is indicated by color (red positive, blue negative) for each 2kb interval across an 80kb window. (D) Density plots showing distribution of epigenetic marks and motifs relative to the center of loop anchors. Normalized tag densities as measured by Homer from H3K27ac ChIP-seq (left) and ATAC-seq (middle) are shown for loops called in iPSC-CM. Grey regions below the peak signals indicate the results from 1,000 null loop sets. CTCF motif frequency per kb (right) is shown for loops called in iPSCs (blue) or iPSC-CMs (red). (E) Network diagram showing two discrete subnetworks of fetal-heart chromatin states at iPSC-CM called loops, with edges connecting statistically significant pairs of chromatin states found at opposing loop anchors. The thickness of the edge indicates the odds ratio of significance, and the presence or absence of an edge indicates statistical significance.

Called chromatin loop sets contain a variety of loop types

To characterize the types of chromatin loops that comprised the loop sets, we examined the distribution of H3K27ac and ATAC peaks, CTCF motifs, and ROADMAP chromatin states from the most epigenetically similar cell type[41] (iPSC for iPSCs; fetal heart for iPSC-CMs) near loop anchors. In both cell types, we found enrichments for active and bivalent chromatin states (Figure 2.2C), H3K27ac (Figure 2.2D left), and chromatin accessibility (Figure 2.2D middle) from their respective cell type above shuffled null loop sets. Additionally, we found that 45.5% of loops had CTCF motifs at both anchors, and that across all loops, CTCF motifs were centrally enriched at anchors (Figure 2.2D right). As seen in Rao et. al[3], the vast majority of loops (85.3%) with CTCF motifs at both anchors had inward facing CTCF motifs. Further, 63.3% and 65.3% loops in iPSC and iPSC-CMs, respectively, were within 25kb of a CTCF ChIA-PET interaction from GM12878[16]. We next examined the types of chromatin states that were statistically significantly paired together (Fisher's Exact $p < 0.05$) and found two subnetworks, one with active chromatin states and the other with repressed or bivalent chromatin, which were discrete in iPSC-CMs (Figure 2.2E) and crossed over through the bivalent states in iPSCs. This crossover, which was only present in iPSCs, is consistent with the role of bivalent and polycomb chromatin in pluripotency[46-48], the role of bivalency in maintaining stem cell region connectivity[48], and with the shift of active states to bivalent and polycomb during differentiation and chromatin rewiring[49]. This result suggests that these specialized roles of bivalent and polycomb chromatin extend to the fine-scale aspects of chromatin architecture, including loops. We next examined the consistency of these loops with previously identified promoter loops from promoter capture Hi-C (pHiC) and found 28.7% and 33.5% of iPSC and iPSC-CM loops to be within 25kb of a

pHiC interaction in these cell types, respectively. Together, these results indicate that the identified chromatin loops include those with active regulatory interactions (e.g. promoter-enhancer interactions), those with repressive interactions (e.g. polycomb complexes), structural loops (CTCF-CTCF), and those with a variety of other types of chromatin states (that were not significantly enriched for being paired together) at their anchors.

Quantification of differential looping between cell types

Statistical methods for finding differential loops across conditions remains a largely open question in the field of chromatin architecture[44]. We found a large proportion of loops which were differentially called, but visually appeared to consistently form across cell types (Figure 2.3A). Thus, to determine if the chromatin loops called in only one of the cell types specifically formed within that cell type, or whether they were also present in the other cell type but not called for technical reasons, we performed a quantitative comparison of the subjects' contact frequencies between the iPSC and iPSC-CM using edgeR[50-52]. For all loops, identified in either one or both cell types, we first compared the total normalized contact frequency (log_2 counts per million, logCPM, obtained via edgeR) of the interactions between both cell types. We observed that the majority of loops that were called in both cell types (grey in Figure 3B left) had high logCPMs in both cell types, whereas the loops that were only called in a single cell type (blue or red in Figure 3B left) tended to have overall low logCPMs and often showed highly similar contact intensities between cell types. We did not observe, however, loops with a high logCPM in one cell type, and a very low logCPM in the other. These patterns

were similar within subjects, suggesting that these subtle modulations in logCPM across cell types were not due to the combination of data across individuals. These results indicate that chromatin loops that were called as differentially present or absent between cell types were often of low logCPM, and were therefore likely to be inconsistently identified by the loop calling algorithms. Thus, the differences in the loop sets between the two cell types were not due to the establishment of novel loops present in only one cell type. We therefore identified loops that showed quantitative differences between iPSCs and iPSC-CMs by statistically comparing normalized read counts across cell types at each loop identified in either cell type (edgeR glmQLFit on Trimmed Mean of M values, TMMs, $q < 0.01$). These cell type-associated loops (CTALs) were identified across a range of logCPM levels and were distinct from those called within each cell type (Figure 2.3B right). This analysis resulted in four loop sets: 1) all loops called in any cell type (union loop set, total: 26,679), 2) loops with statistically higher contact frequency in iPSCs (iPSC cell type associated loops; iPSC-CTALs, total 2,906), 3) loops with statistically higher contact frequency in iPSC-CMs (CM-CTALs, total 2,915), and 4) loops that were not statistically significantly different between the two cell types (non-CTALs, total 20,858). To determine whether 3D architecture at a compartment level contributed to these differences, we identified A and B compartments[3] and partitioned the loops by their location in both cell types. While we found increased contact propensity within A compartments relative to B compartments in both cell types, the percent of variance in logCPM explained by compartment differences was only 0.009. Additionally, we found that the CTAL distribution was consistent across all types of anchor-compartment-cell type combinations. These results suggest that compartment differences

did not drive CTALs. Overall, these analyses establish cell type associated loop sets for future analyses.

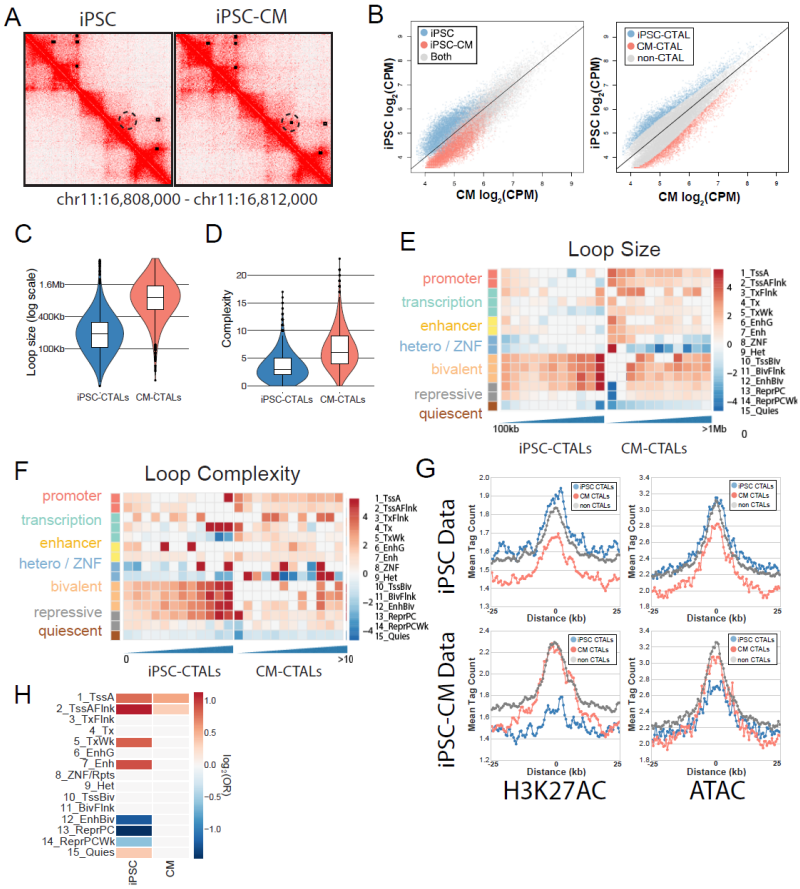

Figure 2.3 Differential chromatin states and sizes in CTALs recapitulate changes in looping across differentiation

(A) Example contact maps showing a loop which appeared in both cell types but was only called in one cell type. Loop calls are shown in the top right half of the contact map as black rectangles. A dotted circle has been added to highlight the region which appears the same in both cell types, but only has a loop call within iPSC-CMs. (B) Scatterplots showing contact frequency in counts per million (CPM) of all loops identified in either iPSCs or iPSC-CMs. The solid black lines indicate the function y = x. (left) Points are colored to indicate loops called in only iPSC (blue), or iPSC-CM (red), or both (gray). (right) Points are colored to indicate loops with significantly increased contact frequency in iPSCs (iPSC-CTAL; blue), iPSC-CMs (CM-CTAL; red), or neither (non-CTAL; gray). (C-D) Violin plots (all four quartiles shown via lower whisker, lower half of box, upper half of box, and upper whisker; lines indicate median) showing distributions of loop size (C), and loop complexity (D) for CTALs. (E-F) Heatmap showing enrichment of regulatory regions near iPSC-CTAL (left) and CM-CTAL (right) at loop anchor centers with loops stratified by (E) size or (F) complexity. The 15 ROADMAP chromatin states of iPSC (E020) or fetal heart tissue (E083) were used, and the $\log_2$ odds ratio of enrichment is indicated by color (red positive, blue negative). CTALs broken down by size into 100kb windows (E), or complexity (F). (G) Line plots showing the mean tag intensity of H3K27AC (left) or ATAC-seq data (left) from iPSC (top) or iPSC-CMs (bottom) at iPSC-CTAL (blue), CM-CTAL (red), or non-CTAL (grey) anchors. Cell type data is enriched at cell type CTAL anchors, and non-CTAL anchors, whereas non-cell type data is depleted at cell type CTAL anchors. (H) Heatmap of $\log_2$(odds ratio) from a Fisher's exact tests for enrichments of differential chromatin states across CTAL anchors. White cells indicate a non-significant Fisher's Exact test (FDR q > 0.05). $\log_2$(OR) is shown by color (red positive, blue negative)

CTALs are associated with differentiation regulatory changes

Previous studies which qualitatively identified cell type specific loops have reported that chromatin architecture becomes more specialized and cell type specific during development[4,45,53]. We examined the physical and regulatory characteristics of iPSC-CTALs and CM-CTALs to determine if these quantitatively identified loops recapitulated these same properties. We observed that CM-CTALs were overall significantly larger (Mann-Whitney $p < 2.2x10^{-16}$; Figure 3C) and more complex (ie shared more anchors with one-another; Mann-Whitney $p < 2.2x10^{-16}$; Figure 3D) than iPSC-CTALs. Additionally, we found active chromatin states to be preferentially enriched at smaller (Figure 2.3E) and less complex (Figure 2.3F) loops. We examined how the enrichment of H3K27ac and ATAC-seq signals varied by CTAL status, and found that within each cell type, CTALs of that cell type and non-CTALs had the highest H3K27AC and ATAC-seq signal, while CTALs of the other cell type were least enriched (Figure 2.3G). These enrichments suggest that loops with decreased contact propensity may be less likely to be involved in gene regulation despite being present in the cell. Next, we examined whether CTALs for each cell type were more likely to overlap cell type specific, or cell type shared, regulatory regions. We found iPSC-CTAL and CM-CTAL anchors to be enriched for differential active promoters, and iPSC-CTAL anchors to be enriched for differential active enhancers (Figure 2.3H, red). These enrichments suggest that CTALs capture cell type specific chromatin dynamics, and are consistent with active elements shifting to repressed elements during differentiation and chromatin rewiring[49] (as enhancers from fetal heart tended to be present in both cell types, but

enhancers in iPSCs tended to be iPSC specific). We also observed that iPSC-CTAL anchors which overlapped iPSC bivalent enhancers were more likely to overlap fetal heart bivalent enhancers (Figure 2.3H, blue), but not the converse, consistent with the repression of active regions of loops during differentiation, and specific use of bivalent chromatin in iPSCs[46,47,49]. Overall, these findings show that CTALs were enriched for cell type specific functional and regulatory regions.

Functional characterization of CTALs

To analyze whether CTALs recapitulated the functional differences between qualitatively identified cell type specific loops, we examined the relationship between contact propensity and eQTLs, differential gene expression, and differential epigenetics across cell types. We first examined whether loops which colocalize iPSC-eQTLs (previously identified from a cohort including these individuals[42]) to the genes that they were statistically associated with (eGenes) had stronger contact intensities within iPSCs than iPSC-CMs. We found a strong enrichment (Mann Whitney-U $p \sim 1x10^{-293}$) for increased iPSC:iPSC-CM contact frequency ratio above non eQTL-eGene loops (Figure 2.4A), indicating that loops with higher contact propensity in a cell type may be more likely to harbor functional genetic variation. Next, we examined whether differential molecular phenotypes were preferentially located at CTAL anchors. We identified differential H3K27ac peaks and genes using ChIP-seq and RNA-seq data generated from iPSC and iPSC-CM samples from the same seven individuals (see methods). We obtained a total of 23,570 differential H3K27ac peaks (DE peaks) and 5,307 differential genes (DE genes) between iPSCs and iPSC-CMs. We found that DE genes and DE peaks

were preferentially located at CTAL anchors (Fisher's exact $p < 0.05$, Figure 4B) compared to the union loop set. Together, these results show that CTALs (loops with quantitative differences in contact propensity across cell types) are associated with cell type specific functions, consistent with previous reports that used qualitatively identified cell type specific loops.

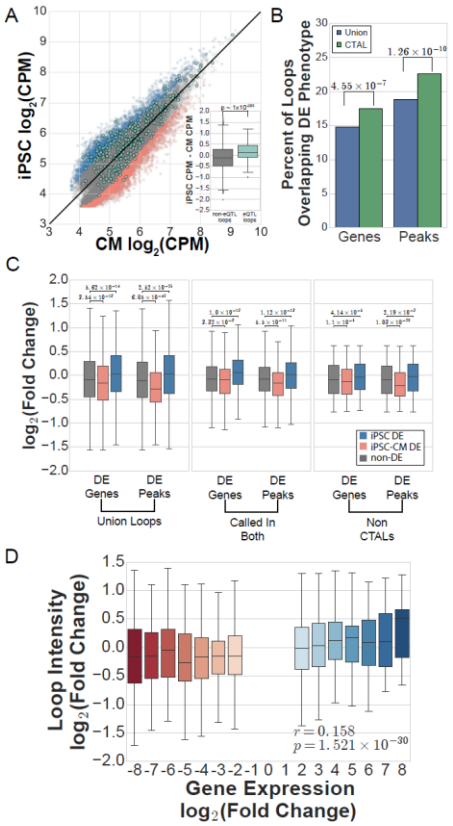

Figure 2.4 Quantitative variation in chromatin loops is associated with differential gene expression and H3K27ac across cell types

(A) Scatterplot showing iPSC vs. iPSC-CM contact frequencies in counts per million (CPM) for all union loops. The black line indicates the y = x function. Background points indicate iPSC-CTALs (blue), CM-CTALs (red), and non-CTALs (grey). Overlaid on this are points indicating iPSC eQTL-eGene containing loops (teal). The boxplot in the lower right corner of the scatter plot shows the fold change between iPSC and iPSC-CM CPMs at non-eQTL loops (grey) or eQTL loops (teal). Positive values indicate a loop had higher CPM in iPSCs, and negative values indicate a loop had higher CPM in iPSC-CM. The p-value was calculated from a Mann-Whitney U test. (B) Barplot showing the percent of CTALs (green) or union loops (blue) which overlap differentially expressed genes or H3K27ac peaks. P-values were found via a Fisher's Exact test on the underlying counts of differentially expressed genes or peaks between union loops and CTALs. (C) Boxplots (all four quartiles shown via lower whisker, lower half of box, upper half of box, and upper whisker; lines indicate median; outliers not shown) of the $\log_2$(fold change) of contact frequency at chromatin loops, with positive indicating higher contact propensity in iPSCs and negative indicating higher contact propensity in iPSC-CMs, for all loops (i), loops called in both cell types (ii), or non-CTALs (iii) with anchors overlapping differentially expressed genes or H3K27ac peaks with higher expression or counts in iPSCs (blue), higher expression or counts in iPSC-CMs (red), or not overlapping a DE gene or peak (grey). P-values were calculated via a Mann-Whitney U test. (D) Boxplot (all four quartiles shown via lower whisker, lower half of box, upper half of box, and upper whisker; lines indicate median; outliers not shown) showing the $\log_2$(fold change) of chromatin loop frequency for chromatin loops overlapping a differentially expressed gene, binned by the $\log_2$(fold change) of the gene. For both expression and chromatin looping, positive indicates stronger counts in iPSCs, and negative indicates stronger counts in iPSC-CMs. The Pearson correlation and p-value shown were calculated on the raw underlying data.

Subtle looping changes are associated with gene regulation

Next, we examined the quantitative association between contact propensity and differential expression, as well as the quantitative association between contact propensity and differential H3K27ac, across cell types. We tested whether the fold change in contact frequency across cell types was in the direction of the cell type with higher differential expression or H3K27ac. We found that across the union loop set, anchors overlapping DE genes with higher expression in iPSCs had significantly greater contact frequency in iPSCs, and anchors overlapping DE with higher expression in iPSC-CMs had significantly higher iPSC-CM contact frequency; similar patterns were found for DE H3K27ac peaks (Mann-Whitney-U $p < 0.05$; Figure 4C left). To establish that this association was due to differences in contact propensity, rather than driven by loops that were differentially called between the two cell types, we examined whether this association was still present within only the loops that were called in both cell types (ie the intersection of iPSC-CM and iPSC called loops). We found that the statistically increased contact frequency (Mann-Whitney-U $p < 0.05$) in the upregulated cell type remained within this set of loops, though the extent of the differences in chromatin looping were smaller (Figure 2.4C middle). Thus, we next examined whether these differences could be observed at non-CTALs (ie loops with non-significant differences across cell types) and found that these loops were still significantly stronger in the expected direction when they overlapped a DE molecular phenotype at their anchor (Figure 2.4C right). These results suggest that subtle variation in chromatin looping across cell types may be functional. Finally, to examine whether chromatin loop contact

propensity proportionally varied with the strength of gene expression differences between cell types, we examined the correlation between fold changes in gene expression and chromatin loop contract frequency at loops with anchors overlapping promoters of differentially expressed genes (Figure 2.4D). We observed a significant correlation ($r = 0.158$, $p < 1.6 \times 10^{-30}$) between the two phenotypes; however, the magnitudes at which the phenotypes varied were quite different, with gene expression varying up to 250-fold, and the middle 3 quantiles of chromatin looping varying less than 3-fold. For these analyses, we pooled data across the genome to measure the association between contact frequency and gene expression, independent of a particular locus; therefore, this analysis compares the relationship between contact propensity and gene expression in aggregate across the genome. As each pair of fold change measurements between contact frequency and gene expression are from the same locus in two different cell types, locus specific biases based on the linear genome which affect Hi-C read depth (number of restriction enzyme sites near the anchors, anchor GC content, and mapping uniqueness)[54] are held constant. Overall, these results suggest that small magnitude changes in contact propensity may be functional as they are associated with large magnitude changes in gene expression across cell types.

Haplotype-based interrogation of loops and gene regulation

To enable the functional characterization of haplotype-specific chromatin looping, we phased the Hi-C, H3K27ac, and RNA-seq data to obtain haplotype-associated phenotype data. We first phased the WGS genotype data for these seven individuals using a combination of Hi-C-based phasing and family structure, resulting in

an average of 2.01M phased heterozygous variants per individual. Next, we assigned informative reads from H3K27ac and RNA expression to each individual's maternal or paternal haplotype using MBASED[55], and then identified significant peaks or genes with allele specific effects (ASE; FDR $q < 0.05$) within each individual using a binomial test. We identified a total of 189 ASE peaks (mean 43 per individual) in iPSCs and 618 ASE peaks (mean 119 per individual) in iPSC-CMs, and 2,582 ASE genes (mean 647 per individual) in iPSCs and 2,214 ASE genes (mean 503 per individual) in iPSC-CMs.

To characterize haplotype-specific chromatin looping, we performed a genome wide analysis to identify haplotype associated chromatin loops with consistent significant allelic imbalance (haplotype associated loops; HTALs) across individuals. Within each cell type, for each individual, we assigned informative Hi-C contacts carrying a phased allele to each haplotype (Figure 2.5A) and examined allelic imbalance across all loops. Next, for each individual, we identified imbalance via a Z score using a half normal distribution (as well as using the computational framework WASP; see Methods for details of complementary analysis), following which we combined the p-values across individuals with Fisher's method for meta-analysis. This process identified 54 total HTALs: 27 from iPSCs, and 27 from iPSC-CMs. We first examined whether these 54 HTALs were enriched for being CTALs of either cell type and found no significant enrichments (Fisher Exact $p > 0.05$). We next examined whether the HTALs were truly cell type specific or if the sparsity of the Hi-C data statistically limited our ability to detect allelic loop imbalance present in both cell types. For each of 7 individuals, we determined the individual's maternal allele ratio for each of the 27 iPSC HTALs using

the iPSC Hi-C data, as well as the maternal allele ratio using the iPSC-CM Hi-C data (Figure 2.5B). We then repeated this process at each of the 27 iPSC-CM HTALs (Figure 2.5C). For both cell types, we found the maternal allele ratios to be highly correlated with the other cell type across all individuals ($0.73 <$ Pearson's $r < 0.97$), which suggests that loop imbalance was consistent across both cell types. As we observed that the maternal allele frequencies were highly correlated across cell types, to increase power for these analyses, for each of the 26,679 chromatin loops in the union set, we pooled contacts for each individual across their corresponding iPSCs and iPSC-CMs. We observed a median of 50 informative contacts per individual per loop, which corresponds to 100% power to identify HTALs with an allelic imbalance ratio of 70% or higher with $\alpha = 0.02$ in an individual, or at $\alpha = 2x10^{-5}$ when all samples display similar imbalance and are combined with Fisher's method meta-analysis. Within each subject, a mean of 6.08% of all chromatin loops showed significant imbalance at $p < 0.05$ (Z score on a half normal distribution; see Methods), slightly higher than the statistically expected 5% by chance; however, only a mean of 0.1% (26.6) were significant under FDR $q < 0.05$ in each individual (Figure 2.5D). To identify HTALs which were consistently imbalanced across individuals, we again combined associations using a Fisher's method meta-analysis for each loop, and identified 7.49% of chromatin loops as HTALs at $p < 0.05$, indicating that consistent allelic imbalance occurs more frequently than by chance. However, only 114 HTALs were significant after multiple testing corrections at FDR $q < 0.05$ (equivalent to $p < 2x10^{-5}$), showing that even with the increased power by using the combined cell type data, the majority of loops had small allelic differences (Figure 2.5E). In comparison, we observed slightly fewer HTALs (N = 89) with the WASP analysis; however, the majority

(83/89, 93%) were found in both sets. These results and power indicate that while we may not detect all small haplotype differences (ie those with imbalance < 70%), large haplotype differences in chromatin looping occur infrequently.

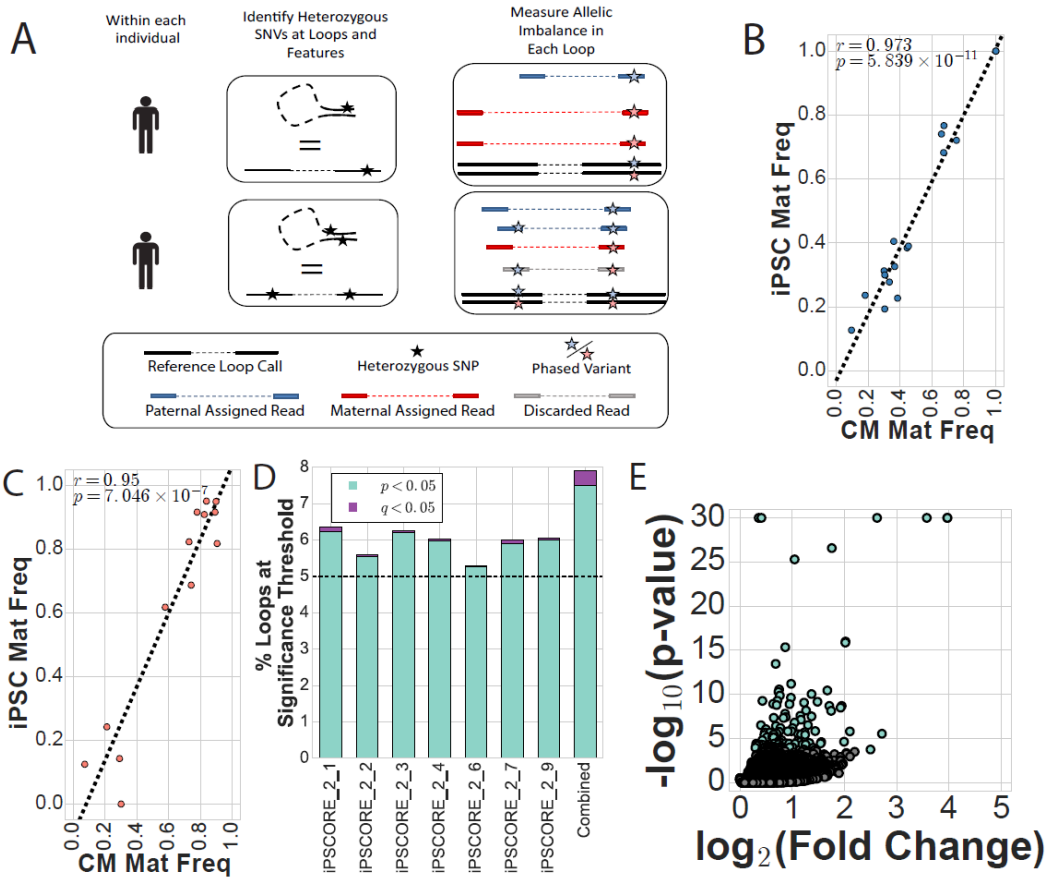

Figure 2.5 Identification of haplotypic differences of chromatin conformation
(A) Schematic showing approach to quantify chromatin loop imbalance within each individual. Examples for two different individuals are shown. Variants were phased using Hi-C and family structure (see methods), and each contact was assigned to its corresponding haplotype based on the phase of heterozygous SNVs it contained. Reference loop calls and unphased heterozygous SNPs are shown in black. Phased variants are shown in red for maternal, and blue for paternal. Reads only overlapping paternal phased variants were assigned to the paternal haplotype (blue reads), and read only overlapping maternal phased variants were assigned to the maternal haplotype (red reads). If a read overlapped both types of variants, it was discarded (grey read). (B-C) Scatter plot showing comparison between iPSC and iPSC-CM maternal haplotype frequencies for one of the seven individuals at HTALs identified in either (B) iPSCs or (C) iPSC-CMs. Linear regression correlation and p-value are reported for each cell type. (D) Barplot showing the percent of loops associated with haplotype imbalance at $p < 0.05$ shown in teal, with those also $q < 0.05$ shown in purple. Bars are shown for each individual separately, or for the results of a Fisher's method meta-analysis p-value (combined; right most bar). A dashed line is drawn at 5% to indicate the number of HTALs expected by chance to be significant at $p < 0.05$. (E) Volcano plot showing the $\log_{10}$(p-value) vs the $\log_2$(fold change of allelic imbalance ratio, with the higher frequency allele always in the numerator of the ratio; see methods) for each loop with the combined data. As only fold changes of major allele frequencies can be calculated due to haplotypes not having a single reference or alternate allele, all fold changes are positive. Significant points (HTALs) are shown in teal.

HTALs are associated with imprinting and CNVs

We next examined whether the 114 genome-wide significant HTALs were statistically more likely to be a specific type of loop, or overlap genomic features previously shown to be associated with differential chromatin looping (imprinted genes[3,16] and somatic and inherited CNVs[30,56]). We first hypothesized that chromatin loops that were variable across cell types may be more variable in general, and thus HTALs would be more likely to be CTALs. We compared the proportion of the 114 HTALs that were also iPSC-CTALs, CM-CTALs, iPSC called, or iPSC-CM called loops to the corresponding proportion of union loops. However, we found no significant differences for any association ($p > 0.05$ for all tests; Figure 2.6A). We next examined whether a particular type of loop was enriched within HTALs (ie CTCF loops, promoter-enhancer loops; see Methods), and found no significant enrichment (FDR $q > 0.05$); together, these results indicate that loops which varied between haplotypes were not more likely to be a specific type of loop. We next compared the distribution of genomic features known to cause large allelic differences within HTALs and the union loop set (Figure 2.6B). We observed that, compared to the union loop set, HTALs were statistically more likely to contain imprinted genes (HTAL: 10.5%; all: 2.7%; Fisher's exact $p = 5.8 \times 10^{-5}$), and somatic (HTAL: 7.0%, all: 1.0%; Fisher's exact $p = 1.8 \times 10^{-5}$) and inherited (HTAL: 27.2%, all: 18.3%; Fisher's exact $p = 2.03 \times 10^{-2}$) CNVs previously identified in these samples[42]. To examine whether these trends held across all levels of imbalance significance, we quantified the extent of association of each genomic feature with chromatin loop allelic imbalance as a function of HTAL p-value. For imprinted genes, as the p-value threshold increased, the odds ratio increased almost log-linearly,

whereas CNV overlap increased but to a lesser extent (Figure 2.6C). We next examined the distribution of the types of CNVs contained within loops by examining the subset of loops which contained any number of only a single type of CNV (Deletion or Duplication, Figure 6D). While we found deletions to be enriched above duplications within union loops (Binomial $p=2.41x10^{-257}$), we found no significant enrichment within HTALs. Thus, while CNV type was not associated with allelic imbalance, loop detection may be affected by CNV presence. The observed pattern of enrichment in deletions is consistent with linearly closer loci having increased Hi-C contact propensity (as deletions reduce the linear space between loci) thereby increasing contact frequency and loop detection power; conversely, duplications increase linear distance and thus decrease contact frequency and loop detection power. Thus, it is unclear how much of this enrichment is due to a technical artifact induced by increased power at deletions. Overall, these results confirm previous reports which suggested that genetic imprinting[3,16] may be a strong driver allelic imbalance, and suggest that CNVs may have smaller effects on allelic imbalance in chromatin looping.

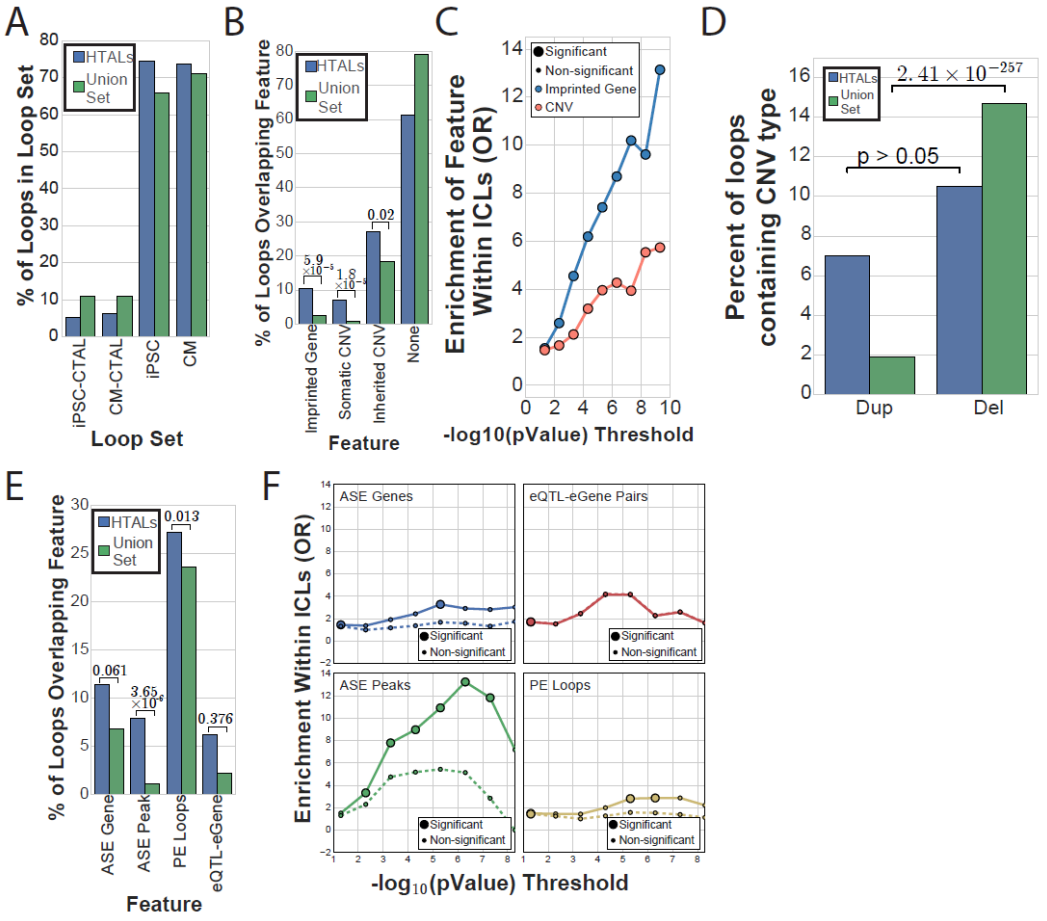

Figure 2.6 Functional characterization of haplotypic differences in chromatin conformation
(A) Barplot showing the percent of union loops (green) or HTALs (blue) contained within each loop-set. (B) Barplot showing the percent of union loops (green) or HTALs (blue) containing the given genomic feature within it (ie the genomic feature overlapped the region between the start of the first anchor and the end of the second anchor). P-values were calculated using via a Fisher's exact test. (C) Line plot showing odds ratio from a Fisher's exact test for HTAL enrichment above the union set for containing an imprinted gene (blue) or containing either an inherited or somatic CNV (red) as a function of the $-\log_{10}$ of the HTAL imbalance p-value. Large circles indicate that the test was significant after Bonferroni correction, and small circles indicate a non-significant association. (D) Barplot showing the percentage of union loops (green) or HTALs (blue) containing only deletions or only duplications. P-values were calculated using a binomial approximation to a normal distribution, adjusted for the number of identified CNVs which were deletions vs duplications. (F) Barplot showing the percent of union loops (green) or HTALs (blue) overlapping the given genomic feature at an anchor. P-values were calculated using a Fisher's exact test. (G) Line plot showing odds ratio from a Fisher's exact test for HTAL enrichment above the union set for containing the labelled feature as a function of the $-\log_{10}$ of the HTAL imbalance p-value, for either all loops (solid lines), or loops that do not contain an imprinted gene or CNV (dashed lines). Large circles indicate that the test was significant after Bonferroni correction, and small circles indicate a non-significant association.

Regulatory genetic variants and contact propensity

We next examined whether HTALs were enriched for functional allele-specific differences by quantifying the enrichment for containing an ASE gene or ASE H3K27ac peak at their anchors, or for being a promoter-enhancer (PE) or eQTL-eGene loop. We found ASE peaks to be enriched at HTAL anchors, and also being a PE loop to be enriched (Fisher's Exact $p < 0.05$; Figure 6E). Notably, despite the increased percentage of eQTL-eGene loops in HTALs, as only 7 eQTL-eGene loops were HTALs (585 eQTL-eGene loops in total), this increase was non-significant. To determine whether regulatory genetic variation was associated with these differences, we excluded the effects from imprinting and CNVs, and examined these associations across a range of imbalance thresholds (Figure 2.6F). The removal of imprinted regions and CNVs greatly attenuated the association, and resulted in a loss of significance for the two molecular phenotypes and PE loop status over almost all ranges of imbalance significance. These results suggest chromatin loops vary across haplotypes much more subtly (ie allelic ratio <70%) than gene expression or H3K27ac, and where variation is larger, it is mainly driven by imprinting and/or CNVs. Additionally, these results show that large allelic imbalances in chromatin loops are primarily restricted to those located in imprinted regions or associated with copy number variation, and that regulatory genetic variants are not associated with large changes in contact propensity.

Haplotypes, contact propensity, and gene regulation

As we observed that subtle differences in contact propensity were quantitatively associated with large differential regulation of gene expression across cell types (Figure

2.4D), we investigated if similar small-scale changes in contact propensity across haplotypes were associated with gene expression and regulation differences. We first compared the general variability of chromatin loops (excluding imprinted regions and CNVs) across cell types (Figure 2.7A) to the variability across haplotypes (Figure 2.7B). We found that more chromatin loops varied to a larger degree across cell types than across haplotypes: ~35% of loops exhibited a $\log_2$ fold change of 0.5 (1.4-fold) or higher across cell types, whereas only ~5% of loops showed a similar fold change across haplotypes (Figure 2.7C). This result suggests that haplotype associated differences are considerably smaller than cell type associated differences. We therefore examined whether the association between contact propensity and gene expression, or contact propensity and H3K27ac, was significant and proportionally consistent across cell types and haplotypes. Across cell types and haplotypes, we found a positive and highly significant correlation (Pearson Correlation; Cell Type: $p = 2.36 \times 10^{-30}$, Haplotype $p = 6.76 \times 10^{-4}$, Figure 2.7D) between gene expression fold change and chromatin loop fold change, and between H3K27ac fold change and loop fold change (Pearson Correlation; Cell Type: $p=6.6\times10^{-211}$; Haplotype: $p=4.63\times10^{-5}$, Figure 7E). Similar to the cell type analyses (Figure 2.4D), we found the range at which gene expression and H3K27ac fold changes occurred to be larger than the range at which loop fold changes occurred. These consistent associations between the cell type and haplotypes analyses, as well as the consistent magnitude differences between looping and molecular phenotype, suggest that large differences in gene expression and H3K27ac are associated with small differences in chromatin loop contact propensity. Additionally, as the association between gene expression and contact propensity was consistent across haplotypes, these results suggest

that genetic variation could exert effects on gene expression through small modulation of contact propensity.

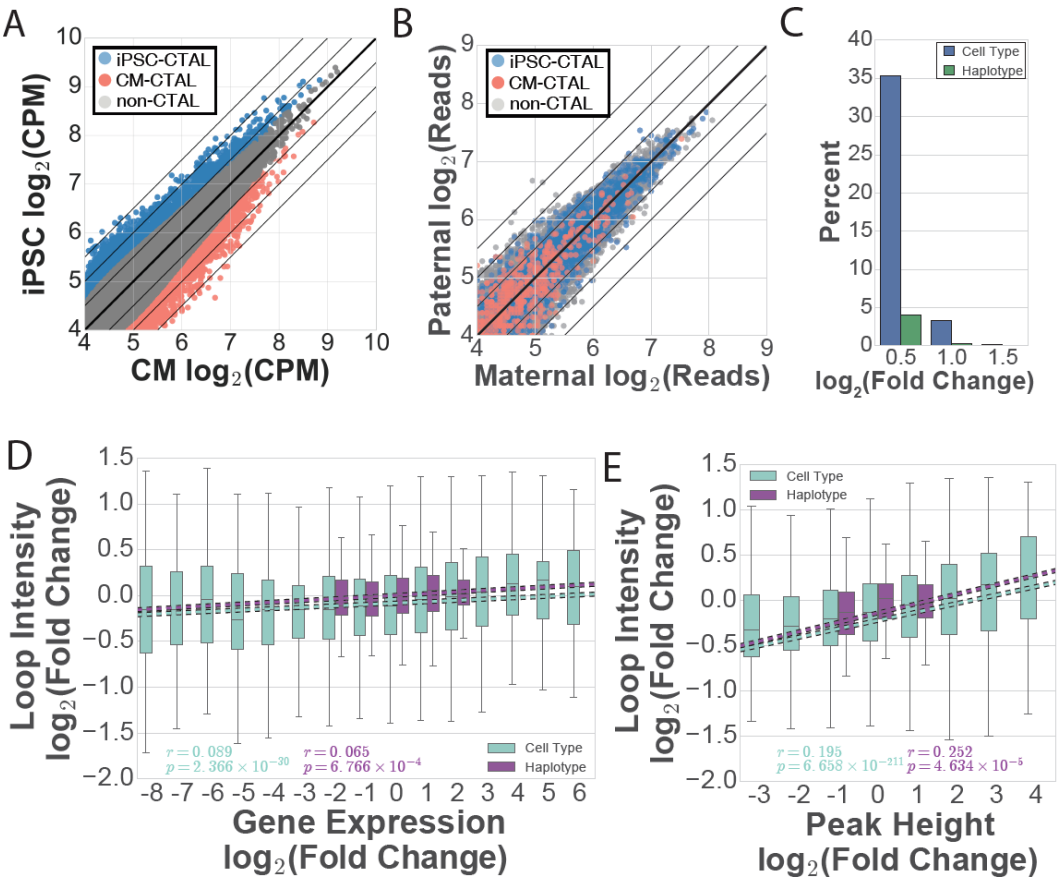

Figure 2.7 Comparison of chromatin loop, gene expression, and H3K27ac variability across cell types and haplotypes

(A-B) Scatterplots showing (A) contact frequency in $\log_2$ counts per million ($\log_2$(CPM)) across cell types or (B) read counts across haplotypes of all union loops colored by CTAL status. The solid bold lines indicate the function y = x, and other lines indicate absolute fold changes of $\log_2(0.5)$, $\log_2(1)$, and $\log_2(1.5)$. (C) Percent of loops with at least the shown $\log_2$(Fold Change) or across cell types (blue) or haplotypes (green). (D-E) Boxplot (all four quartiles shown via lower whisker, lower half of box, upper half of box, and upper whisker; lines indicate median; outliers not shown) showing the $\log_2$(fold change) of chromatin loop contact frequency for chromatin loops overlapping a (D) differentially expressed or ASE gene, or (E) differential or ASE H3K27ac peak, binned by the $\log_2$(fold change) of the (D) gene or (E) peak. Boxes are shown for cell type comparisons in teal, and haplotype comparisons in purple, linear regressions are plotted with dashed lines, and r's and p-values are shown and colored from the raw data in each data set independently. For all data, positive fold change indicates stronger counts in iPSCs, and negative fold change indicates stronger counts in iPSC-CMs.

Discussion

Here, we generate a resource of phased genotypes, Hi-C, and molecular phenotype data in two cell types for seven individuals who are a part of a three-generation family, and use this data to perform an in depth, genome-wide, functional examination of changes in contact propensity across cell types and haplotypes. These chromosome-length haplotypes, and accompanying phased data, will enable future studies examining long range interactions between multiple genetic variants on the same chromosome. Additionally, these Hi-C maps are the highest resolution maps for human iPSCs and iPSC-CMs currently available and are thus an important resource for the prioritization of functional variants and their potential gene targets in these cell types.

We performed quantitative comparisons of contact frequency across cell types and haplotypes to identify differences in chromatin looping, and integrated these differences with quantitative measures of differential expression and H3K27ac to examine the functionality of contact propensity. These analyses revealed a proportional association between contact frequency and gene expression/H3K27ac, which surprisingly linked the phenotypes across different magnitudes of variability: extremely subtle changes in contact frequency were associated with large differences in gene expression and H3K27ac. If contact propensity at loops is a fundamental regulator of gene expression, differences in contact propensity would be expected to be associated with similarly sized differences in gene expression regardless of the environment in which the differences occurred (ie across cell type, haplotype, or experimental conditions). As we observed a consistent relationship between the two, we believe these data indicate that

contact propensity is a mechanism involved in regulating gene expression, similar to enhancer activity or transcription factor binding strength. Notably, as we identified a non-directional correlation, contact propensity may either affect, or be affected by, gene expression and/or regulation.

While the mechanisms underlying changes in contact propensity are currently unknown, there are several reasonable hypotheses. Previous studies showing that the physical 3D structure of the genome can be reconstructed from contact frequency via polymer physics models[36,57-60] suggest that contact propensity could result from changes in spatial proximity. The fact that CTCF and Pol2 ChIA-PET show similar profiles to Hi-C data[16] suggest that differences in protein binding near loop loci could also affect contact propensity. Finally, as we found associations between contact propensity and H3K27ac, regulatory chromatin activity could modulate contact propensity. Future studies examining these mechanisms could provide insights into the biological processes underlying differential contact propensity and gene regulation.

The identification of specific causal variants associated with differential contact propensity is likely to be challenging, as we did not find a large number of HTALs with strong effects outside of imprinted and copy number variable regions. As the effects of imprinting are parental in nature, rather than genetic, it is necessary to search outside of these regions for causal regulatory variants. In non-imprinted regions, if we interpolate the association between gene expression and contact propensity, the linear model would suggest that a gene with 98% ASE would be expected to be associated with a loop

imbalance of only ~52%. This minute difference in loop imbalance provides a possible explanation for why we did not observe HTALs associated with gene regulation or ASE, but found a quantitative association between Hi-C signal and functional phenotypes overall. Additionally, it suggests that high coverage would be needed to identify HTALs outside imprinted regions. Thus, for the validation of specific variants, or identifying loop QTLs, future studies should consider using an unbiased targeted loop capture assay with higher sensitivity and targeted coverage than Hi-C, such as sequence-based pHi-C, and perform quantitative analyses using these data.

Finally, our work provides some insight into the ongoing question of whether changes in chromatin looping cause changes in gene expression, or if changes in gene expression cause changes in looping[3,13,37,45,53,61-64]. It has been established that the creation of new chromatin loops can alter gene expression[65], however is has been less clear whether altering gene expression results in meaningful changes in chromatin loops[45,61,66]. Evaluating whether chromatin loop changes are meaningful requires an understanding of the scale at which functional changes in chromatin loops occur. As our findings suggest that subtle changes are functional, we believe these discordant interpretations could have arisen from studies either not being sufficiently powered to detect small effects, or from discounting small changes as nonfunctional. Our work therefore provides a foundation for future studies to quantitatively examine how changes in contact propensity elicit changes in expression (or vice versa) and suggests that studies designed to detect small magnitude changes in chromatin loop variability may be needed to delineate the relationship between chromatin loop imbalance and gene expression.

Author Contributions

Conceptualization, W.W.G., H.L., E.N.S., and K.A.F.; Methodology, W.W.G., E.N.S., and K.A.F.; Software, W.W.G., H.L., and H.M.; Validation, H.L.; Formal Analysis, W.W.G., H.L., P.B., M.D., and D.J.; Investigation, W.W.G., H.L., A.D.A., P.B., A.S., and S.S.; Data Curation, W.W.G., H.M.; Writing – Original Draft, W.W.G., H.L., E.N.S., and K.A.F.; Writing – Review & Editing, W.W.G., E.N.S., and K.A.F.; Visualization, W.W.G., and H.L; Supervision, E.N.S., and K.A.F.; Project Administration, K.A.F.; Funding Acquisition K.A.F.

Acknowledgments

This work was supported in part by a California Institute for Regenerative Medicine (CIRM) grant GC1R-06673 and NIH grants HG008118-01, HL107442-05, DK105541-03 and DK112155-01. RNA-seq were performed at the UCSD IGM Genomics Center with support from NIH grant P30CA023100. WWG was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number F31HL142151. DJ was supported by the National Library of Medicine Training Grants T15LM011271. PB is supported in part by the Swiss National Science Foundation Postdoc Mobility fellowships P2LAP3-155105 and P300PA-167612. Whole genome sequencing was performed at Human Longevity, Inc. Arima Genomics was supported by NIH grant R41HG008118. We would like to acknowledge Yunjiang Qiu for scripts and help with qualitative Hi-C loop calling.

Data Availability

All genomic data are available through dbGAP accessions phs000924 (Hi-C, RNA-seq, CHiP-seq, ATAC-seq) and phs001325 (whole genome sequence SNV and CNV genotypes). Processed data files are available through GEO entry GSE125540

Code for correcting switch errors using family structure is available at https://github.com/billgreenwald/HiC-Family-Phaser

Competing Interests

Drs. Anthony Schmitt and Siddarth Selvaraj are employees and stockholders at Arima Genomics. They generated Hi-C libraries and data, and provided advice on phasing and related analyses, but did not influence the scientific outcome of this work.

Methods

Subject enrollment

The seven individuals used in this study were recruited as part of the iPSCORE project[38]. We have complied with all relevant ethical regulations for work with human participants, and informed consent was obtained. iPSCORE recruitment was approved by the Institutional Review Boards of the University of California, San Diego and The Salk Institute (Project no. 110776ZF), and consent forms were received from each subject. Subject information including sex, age, and ethnicity were collected during recruitment. Skin biopsy was performed to obtain fibroblasts for iPSC reprogramming, and blood samples were collected for whole genome sequencing.

iPSC derivation and iPSC-CM differentiation

Cell line derivation and differentiation were performed as described in Benaglio *et al*[41]. From the seven individuals, fibroblast samples from skin biopsies were reprogrammed using non-integrative Cytotune Sendai virus (Life Technologies)[39] following the manufacturer's protocol. Each independent reprogramming resulted in one or more iPSC clones of the subject. At passages 12-13, genomic integrity of at least one iPSC clone per subject was assessed using Illumina HumanCoreExome arrays, and pluripotency of iPSCs was assessed for most clones in this study by flow cytometry of the pluripotency markers SSEA4 and TRA-1-81[38]. iPSCs of each clone were harvested between passages 12 to 40, resulting in a total of 38 iPSC samples used in this study. Each iPSC clone was then used to generate multiple independent iPSC-CM differentiations using a monolayer protocol[40], resulting in a total of 27 iPSC-CM samples used in this study. Among these iPSC-CM samples, 11 of them were subjected to purification via 4 mM Sodium L-Lactate at Day 15 after the start of differentiation and collected at Day 25[67]; one iPSC-CM sample was subjected to lactate purification at Day 11 and collected at Day 16; the rest of the iPSC-CM samples were not subjected to lactate purification and collected at Day 15. Across all molecular assays detailed below, lactate purified and non-lactate purified iPSC-CM samples showed similar profiles; we therefore combined data across the two protocols. Single-nucleotide variants (SNVs) and copy-number variants (CNVs) of these individuals were obtained from ~40X whole-genome sequencing (WGS) from iPSCORE[38] through dbGAP phs001325.v1.p1 and by DeBoever et. al.[42].

Hi-C data generation

For each of the 11 iPSC and 13 iPSC-CM Hi-C samples, we performed *in situ* Hi-C on 2-5 million cells. Hi-C libraries were prepared using *in situ* Hi-C[3]. Cells were crosslinked at a final concentration of 1% formaldehyde and quenched using 200 mM glycine. Crosslinked cells were then lysed and nuclei were digested with 100U MboI overnight at 37°C. Next, fragmented ends were biotinylated for 90min at 37°C, and the sample was diluted and proximity ligated for 4 hours at room temperature. Crosslinks were reversed by the addition of SDS, ProteinaseK, and NaCl, and allowed to incubate overnight at 68°C. Samples were then purified by ethanol precipitation, resuspended in 100uL 1X Elution Buffer, fragmented using a Covaris S2 instrument, and size selected using AmpureXP beads. Subsequently, biotinylated ligation junctions were pulled down using T1 Streptavidin beads. Hi-C libraries were prepared using streptavidin beads by performing end-repair, dA-tailing, and adapter ligation, following which PCR amplification and purification was performed. The resulting libraries were sequenced on an Illumina HiSeq 4000 machine to obtain 150bp paired-end reads.

RNA-Seq data generation

RNA-seq data was obtained from Benaglio *et. al*[41]. Specifically, total RNA was isolated using the Qiagen RNAeasy Mini Kit from frozen RTL plus pellets, including on-column DNAse treatment step. RNA was eluted in 60 µl RNAse-free water and run on a Bioanalyzer (Agilent) to determine integrity.

Concentration was measured by Nanodrop. Illumina Truseq Stranded mRNA libraries were prepared and sequenced on HiSeq2500, to an average of 40 M 100 bp paired-end reads per sample. RNA-Seq reads were aligned using STAR[68] with a splice junction database built from the Gencode v19 gene annotation[69]. Transcript and gene-based expression values were quantified using the RSEM package (1.2.20)[70] and normalized to transcript per million bp (TPM).

ChIP-Seq data generation and peak calling

H3K27ac data was obtained from Benaglio *et. al*[41]. For H3K27ac, 2 x 10^6 fixed cells were lysed in 60 µl of MAGnify™ Chromatin Immunoprecipitation System Lysis Buffer (Thermo Scientific) and sonicated using Bioruptor 200 (Diagenode) for 35-45 min of 30 sec on/30 sec off cycles. H3K27ac antibodies (Abcam ab4729, lots GR183922-2 (1.75 µg) or GR184333-2 (1 µg)) were coupled for 2 hours to ProteinG Dynabeads (Thermo Scientific), and used for overnight chromatin immunoprecipitation in IP buffer (1% Triton-X, 0.1% DOC, 1x TE, 1x Roche Complete Proteinase Inhibitor tablets (RCPI)). Beads were washed five times with washing buffer (50 mM Hepes pH 8, 1% NP-40, 0.7% DOC, 0.5M LiCl, 1mM EDTA and 1x RCPI) and once with TE buffer. DNA was eluted and reverse crosslinked overnight in elution buffer (10 mM Tris-HCl pH 8, 1 mM EDTA, 1% SDS) at 65˚C. DNA was purified using Qiagen MinElute PCR Purification kit, quantified by Qubit (Thermo Scientific) and submitted to library preparation and barcoding using KAPA Hyper Library preparation kit (KAPA

Biosystems). Libraries were sequenced on an Illumina HiSeq2500 or a HiSeq4000 to an average of 35 M 100 bp paired-end reads per sample.

ChIP-Seq reads were mapped to the hg19 reference using BWA[71]. Duplicate reads, reads mapping to blacklisted regions from ENCODE, reads not mapping to chromosomes chr1-chr22, chrX, chrY, and read-pairs with mapping quality Q <30 were filtered. Peak calling was performed using MACS2[72] ('macs2 callpeak -f BAMPE -g hs -B --SPMR --verbose 3 --cutoff-analysis --call-summits -q 0.01') using pooled BAM files from all iPSC or iPSC-CM samples and with reads derived from sonicated chromatin not subjected to IP (ie input chromatin) from a pool of samples used as a negative control.

ATAC-Seq data generation and peak calling

ATAC-seq data was obtained from Benaglio *et. al*[41]. Specifically, the ATAC-Seq protocol has been adapted from Buenrostro et al.[73]. Frozen nuclear pellets of 5 x 10^4 cells each were thawed on ice, suspended in 50 μL transposition reaction mix (2.5 μL Tn5 transposase in 1x TD buffer, Illumina Cat# FC-121-1030), and incubated for 30 min at 37°C. Reactions were purified using Qiagen MinElute kit, eluted in 10 μL water and amplified using the KAPA real-time library amplification kit (KAPA Biosystems) with barcoded adaptors. PCR reactions were terminated after 10 to 13 cycles and purified using AmPure XP beads (Beckman Coulter). Samples were size selected using SPRIselect beads (Beckman Coulter) to a size range of 150 to 850 kbp and sequenced on an Illumina HiSeq2500 to an average depth of 30 M 100 bp paired end reads.

ATAC-Seq reads were aligned using STAR to hg19 and filtered using the same protocol as for ChIP-Seq. In addition, to restrict the analysis to regions spanning only one nucleosome, we required an insert size no larger than 140 bp, as we observed that this improved sensitivity to call peaks and reduced noise. Peak calling was performed using MACS2 on merged BAM files of iPSC and iPSC-CM meta-samples with the command 'macs2 callpeak --nomodel --nolambda --keep-dup all --call-summits -f BAMPE -g hs', and peaks were filtered by enrichment score ($q < 0.01$).

Creation and analysis of Hi-C contact maps

For each sample, Hi-C reads were first aligned to human reference genome hg19 using BWA-MEM (version 0.7.15)[71] with default parameters. Forward and reverse reads from the paired-end data were aligned independently to allow for identification of split reads that represent ligations between two genomic loci due to spatial proximity[3]. Paired-end reads were then reconstructed, processed, and filtered using the Juicer pipeline[8], resulting in the removal of: unmapped reads, abnormal split reads (split reads that cause ambiguous positioning of the contact), read pairs within the same restriction enzyme fragment, low mapping quality read pairs (MAPQ < 30), and duplicate reads. Subsequently, read pairs that were less than 2kb apart were removed to avoid self-ligated fragments. These filtered read pairs (contacts) were subsequently used to generate chromatin contact maps for each sample via Juicer. To create Hi-C contact maps on a per individual basis, contacts were pooled across all samples of a particular cell type for each individual, and to create maps of iPSC and iPSC-CM, contacts were pooled across individuals within the respective cell type. These processes resulted in a set of binary .hic

files, which were utilized to obtain raw and Knight-Ruiz (KR)[74] normalized counts as well as normalization vectors of contact frequency matrices via Juicebox command line tools[12] at various resolutions used throughout this study.

Correlation of Hi-C contact maps between samples

The KR normalized contact matrices of each sample were retrieved from the .hic files at 1Mb using Juicebox[12]. The contact matrices were then vectorized in order to calculate Pearson correlation between each of the samples in R. Hierarchical clustering analyses of the Pearson correlation were performed in R using hclust with default settings and (1- Pearson correlation) as dissimilarity height. HiCRep was run using the default parameters on chromosome 22 as suggested by the documentation[43].

Identification of chromatin loops

Chromatin loops in iPSC and iPSC-CM were called using both Fit-Hi-C[9] and HICCUPS[3,12]. For Fit-Hi-C, loops were called in meta-fragment resolutions that each contained a fixed number of consecutive restriction enzyme (RE) fragments, ranging from 10 to 30 RE fragments. First, significant interactions (FDR $q < 0.01$) were identified through jointly modeling the contact probability using raw contact frequencies and KR normalization vectors with the Fit-Hi-C algorithm (Step 1). Next, the output of Fit-Hi-C was pruned by requiring that: 1) the interaction itself was significant; and 2) for each anchor of the interaction, 3 of the 5 immediately upstream or downstream bins from the opposing anchor were significant (Step 2). We then merged high-confidence interactions within 20kb using pgltools[75] (Step 3), discarded interactions that did not have any other

interactions within 20kb, and retained the most significant call at each interaction event (Step 4).

For HICCUPS, loops were called using fixed-size bin resolutions from 5kb to 25kb at 1kb bin size intervals. Briefly, default parameters of peak size (p) and window size (i) were used to call loops at 5kb and 10kb resolutions provided by HICCUPS[12], and parameters for other resolutions were chosen by linearly scaling the parameters with respect to the resolution chosen. Specifically, for 6kb, 7kb, 8kb, and 9kb resolutions, the values of these two parameters were interpolated from the 5kb and 10kb values, and rounded to the closest integer. For resolutions greater than 10kb, the default 10kb parameters were used. Following loop calling, as performed by Rao et al.[3], for resolutions from 5kb to 10kb, loops within 20kb were merged using pgltools. For resolutions above 10kb, loops within twice the size of the anchor were merged using pgltools. At each merging event, the loop call with the most statistical significance provided from HICCUPS output was retained.

Loop calling techniques are known to be technically variable[44]. We found many loop calls from both Fit-Hi-C and HICCUPS that were located at random points throughout the Hi-C matrix far off the diagonal. We thus developed a procedure to remove these loop calls by examining the number of resolutions at which the loop was identified. We intersected loop calls across all resolutions within each calling method, retaining the highest-resolution call at each intersection event, and filtered out loops present in less than 3 or 7 resolutions for HICCUPS or Fit-Hi-C, respectively. The loops

retained in these filtered sets visually appeared to best represent the underlying Hi-C data. Next, we compared how these filtered sets overlapped with promoter-capture HiC[76] or the Rao. et al loop set and found that using these filtering criteria resulted in a higher overlap with the retained loops, suggesting that this filtering strategy removed spurious loop calls. After this filtering, while we found a large number of loops that overlapped between Fit-Hi-C and HICCUPS, many loops were unique to only one caller. We therefore intersected the loops across calling methods, retaining the loop with the smallest total anchor size at each intersection event. Overall, this process retained the smallest resolution loop call for all loops present in either 3 HICCUPS or 7 Fit-Hi-C resolutions, and resulted in the iPSC called and iPSC-CM called loop sets.

Identification of TADs

To identify TADs, we utilized the HMM method from Dixon et. al 2012[20] with the Hi-C matrix at 40kb resolution as recommended. To determine the percent of loops that were at TAD boundaries, we paired TAD boundaries sequentially in the file to create a pgl format file, and then used pgltools intersect to find the percent of loops with both anchors at TAD boundaries.

Identification of Compartments

Chromatin compartments were called for each cell type via Juicer command line tools using the corresponding .hic files where the first PC of the normalized contact frequency matrices were extracted at 1Mb resolution. The signs of the PC eigenvectors

were used to stratify each chromosome into two arbitrary compartments. To determine the activity status of the two compartments on each chromosome, we counted the number of reads from 1) RNA-seq, 2) H3K27ac ChIP-seq, and 3) ATAC-seq aligned to each of the 1Mb bins from all available samples for each cell type, averaged the read counts across all samples for each assay in each cell type, and assigned the compartment with higher average read counts from all the three assays as the active compartment (A) and the other compartment as inactive compartment (B). While most of the time all three assays had consistent compartment activity calls, chr21 of iPSC and chr22 of iPSC-CM had inconsistent calls, where we assigned the compartment activity based on the majority of assays.

Creation of the union loop set

To create the union loop set, we used pgltools merge to find all loops from the iPSC call set and iPSC-CM call set with both anchors within 20kb. This process led to merge events of 1, 2, or 3 loops, which were resolved as follows: 1) if there was only 1 loop present within 20kb (ie, only 1 loop set had a call), this loop was retained, 2) if there were 2 loops present within 20kb, the loops were merged by pgltools merge, 3) if there were 3 loops present, pgltools closest was used to identify which two loops were closest together; these two loops were merged, and the third loop was retained as its original call.

Identification of cell type associated loops (CTALs)

After filtering contacts with Juicer, raw contact frequencies for union loops were obtained by intersecting the filtered read pairs from the 11 iPSC and 13 iPSC-CM Hi-C

samples with the union loop set using pgltools coverage. These raw contact frequencies were used as input in edgeR[52], normalized to remove library size bias using trimmed mean of M-values (TMM), and compared between the 11 iPSC and 13 iPSC-CM samples using quasi-likelihood F-test. By comparing Hi-C read coverages at the same genome loci in two cell types, the linear genome biases that are known to affect Hi-C are held constant (restriction enzyme cut site frequency, GC content, and mappability)[54]. The significant differential loops were determined by FDR adjusted $q < 0.01$.

Creation of null loop sets for functional comparisons

As chromatin loops, and genome annotations such as chromatin states, are highly structured and depend on genomic distance both between their own anchors and other chromatin loops, we used permutation to test for functional enrichment within chromatin loops and at loop anchors. We generated 1000 null loop sets for both the iPSC called and iPSC-CM called loop sets to use for statistical analysis, as genome-wide background levels of genomic traits may not accurately represent a true random distribution of paired-genomic loci. The null loops were generated for each chromosome by: 1) removing the gap regions on the human reference genome obtained from UCSC genome browser (https://genome.ucsc.edu/) and updating the loop positions according to this no-gap-genome; 2) sliding the loop positions on the no-gap-genome for a consistent random distance d such that $2Mb < d <$ chromosome size - 2Mb for each null set; and 3) gap regions were added back to the genome, null loop positions were updated back to hg19. In step 2, when loop positions moved beyond the chromosome size after rotation, loops were instead moved to the beginning of the chromosome. Null loops with anchors

overlapping a gap region were removed (an average of 0.5% loops were removed in each cell type).

Distribution of motifs and tag frequencies at anchors

The findMotifsGenome.pl script from HOMER (v4.7) was used to determine enriched motifs at loop anchors, using the entire size of the anchor as the search space. The HOMER script annotatePeaks.pl was used to identify the distribution frequencies of CTCF motifs, H3K27ac ChIP-seq reads, or ATAC-seq reads in each set of loop anchors with a bin size of 500bp and a window size of 50kb using all bam files for the respective molecular phenotype simultaneously.

Determining enrichment of chromatin states at loop anchors

For each of the ROADMAP tissues[77], the core 15-chromatin-state models were obtained as BED format from http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state, and the states were separated into their original 200bp bins. To determine the enrichment of each chromatin state at a loop anchor, we compared the proportion of 200bp bins in the state of interest on the loop anchor, to the genome-wide background level of the bins via Fisher's exact test. A significance level of $p < (0.05 / 15)$ was considered significant.

Identification of differential peaks and genes

To identify differential H3K27ac peaks and genes, we first used featureCounts[78] to obtain the number of reads for each assay from each gene as annotated in gencode v19,

or from each peak identified by merging all the H3K27ac data together. Next, we used DEseq2 v1.10.1[79] with default parameters to identify differential peaks and genes with a $\log_2$(fold change) >2 and an FDR corrected q-value < 0.05.

Enrichment of cell type specific regulatory regions at CTALs

To determine if cell type specific regulatory regions were enriched at CTALs, for each cell type, we first split the union loop set into CTALs and non-CTALs. Next, we examined whether the proportion of CTALs overlapping a cell type specific regulatory region was statistically larger than the proportion of non-CTALs. For example, to test whether iPSC-CTALs were more likely to harbor an iPSC-specific active promoter, we restricted the analysis to loops overlapping an iPSC active promoter, and tested whether the proportion of loops overlapping an iPSC specific active promoter was higher within CTALs than non-CTALs. For all analyses, we used Roadmap E020 (iPSC) for iPSCs, and Roadmap E083 (fetal heart) for iPSC-CMs. We defined an anchor as overlapping a cell type specific regulatory region as an anchor which overlapped the region in the tested cell type (E020 for iPSC-CTALs and E083 for CM-CTALs), but did not overlap the region in the other cell type (E083 for iPSC-CTAL comparisons, E020 for CM-CTAL comparisons).

Phasing genomes

To obtain accurately phased genotypes for each sample, we performed initial phasing using the Hi-C data, and then subsequently utilized family structure to identify, and fix or remove, haplotyping errors (point errors). We first determined the initial

phased genotypes for each individual, at each site at least one individual was heterozygous, by analyzing the HiC data with Haploseq[80]. Next, as Haploseq only identifies heterozygous sites, we filled in missing genotype data with unphased genotypes from iPSCORE WGS variant calls for these individuals. To determine the corresponding parental haplotype for each child haplotype (parent-child haplotype combination), we identified the average concordance between each child haplotype, and each of the four parental haplotypes, in 1MB bins chromosome by chromosome, and identified the best matching parent-child haplotype combination for each child chromosome. Within each parent-child haplotype combination, we identified meiotic recombinations within the parent so that we could identify and fix point errors across the genome. We identified recombinations by finding the extreme points from the following scoring function: for a given child haplotype C1, haplotypes from a single parent PH1 and PH2, and N heterozygotic sites across the genome in the child,

$$Score = \sum_{i=1}^{N} \begin{cases} 1 \; if \; C_i = \; PH1_i \; and \; C_i \neq \; PH2_i \\ -1 \; if \; C_i = \; PH2_i \; and \; C_i \neq \; PH1_i \\ 0 \quad otherwise \end{cases}$$

We then split each parent-child haplotype combination into crossover blocks at each crossover position so that each child SNV could be compared to both matching parental haplotypes simultaneously, and fixed switch errors according to Mendelian inheritance. Additionally, if any member of the family was unphased at the site, we phased these variants to follow Mendelian inheritance, generating switch error free genotypes. After phasing each trio individually, we re-evaluated Mendelian inheritance

across all seven individuals, and removed any sites where Mendelian inheritance was violated, as these indicated genotyping errors in one or more individuals.

Identification of genome-wide imbalanced chromatin loops

To identify Haplotype Associated Chromatin Loops (HTALs), we phased contacts from each chromatin loop in the union loop set across cell types, and identified allelic imbalance that was statistically significant at a genome wide threshold. We first identified all contacts within 25kb of a loop, kept those containing at least one heterozygous SNV, and discarded those with no heterozygous SNVs. Next, using all BAM files for each individual (11 iPSC BAMs across 7 individuals, and 13 iPSC-CM BAMs across 7 individuals), we assigned contacts to their matching haplotype when all heterozygous SNVs matched a single haplotype, and discarded other contacts. We did not remap reads with WASP as 1) the alignment scores from the single end bams do not reflect the true mapping scores of the Hi-C contact due to the highly chimeric nature of Hi-C reads, and 2) the insert size that appears from normal paired-end mapping of Hi-C reads, and thus cannot be filtered by WASP. At each loop, we then calculate a Z score via a binomial approximation to a normal distribution from the greater and lesser allele counts, always using the greater allele as the test variable, and then calculated a p-value from a half-normal distribution for each person to account for the imbalance values being > 0.5 by definition. When comparing Hi-C counts across haplotypes, biases known to affect Hi-C read depth are held constant as the genomic loci are held constant (see CTALs methods for details. To obtain a single p-value for imbalance of each loop, we use Fisher's method to obtain a meta-p-value across all 7 individuals. Finally, to identify

genome-wide significant HTALs, we use the Benjamini-Hochberg FDR correction to obtain a q-value, and identified loops with a q-value < 0.05 as genome-wide significant HTALs.

To identify HTALs with a beta-binomial test, we utilized the combined haplotype scripts from WASP. First, we created a CHT input file using the haplotype counts for each loop. Next, we passed these files to fit_as_coefficients.py to calculate the binomial overdispersion parameters. Finally, we obtained p-values for each individual separately from combined_test.py with the option –as_only. These p-values were combined via Fisher's method and both the combined and raw p-values were used for downstream analyses. This analysis resulted in the identification of 89 HTALs, 83 of which were contained in the half normal HTAL set (93%). We repeated the analyses from Figure 6 using these results and observed similar enrichment patterns to the half-normal approach, but found stronger enrichments at imprinted loci.

Calculation of power to detect HTALs

To determine the power to identify chromatin loop imbalance at different allelic imbalance fractions, we calculated Z scores as above using parameters for numbers of contacts (ranging from 5-100 in steps of 5), allelic imbalance fractions (from 0.55-0.95 in steps of 0.05). We then calculated the power from a half-normal distribution using alpha thresholds ranging from $1x10^{-x}$ to $9x10^{-x}$ for any integer $2 \leq x \leq 6$ within each individual.

We then calculated the alpha threshold from a meta p-value obtained from combining seven individuals displaying the same imbalance via Fisher's method.

Chromatin state enrichments at HTALs

To examine whether any pairs of chromatin states were enriched at opposing HTAL anchors, we annotated all HTAL anchors with the chromatin states they overlapped (with iPSC or fetal-heart chromatin states) via pgltools intersect1D. Next, we used a Fisher's Exact test for each pair of states (125 pairs total) to compare the proportion of HTALs with the states at their anchors to the proportion of non-HTALs. Finally, to correct for multiple testing, we performed FDR correction on the p-values.

Loop set enrichments at HTALs

To examine whether any loop sets (CTALs, cell type called loops, CTCF ChIA-PET interactions, or pHiC interactions) were enriched for HTALs, we annotated HTALs by the loops they overlapped via pgltools intersect. Next, we used a Fisher's Exact test for each loop set to test for enrichment of the loop set within HTALs relative to non-HTALs.

ASE gene and peak identification

To identify genes and peaks exhibiting genome wide significant allele-specific expression (ASE) from RNA-seq or ChIP-seq data, within each cell type, for each individual, we pooled all samples by cell type, applied WASP[81] to reduce reference allele mapping bias, used MBASED[55] (R package version 1.4.0) to obtain allelic ratios and p-

values for each gene and peak for each individual, and identified significant genes or peaks as those with an FDR corrected q-value < 0.05.

Chromatin loop set and genomic feature enrichment for HTALs

To identify chromatin loops containing imprinted genes or CNVs, we utilized the pgltools findLoops function to create a bed file from the union loop set, and then used bedtools[5] intersect function to obtain all loops containing the genomic characteristic. To identify ASE genes overlapping chromatin loop anchors, we utilized pgltools intersect1D function. To identify eQTLs polymorphic in the family with eGenes connected by a chromatin loop, we created a set of all eQTL-eGene pairs with empirical p < 0.05 from DeBoever *et al.*[42] in the PGL format, and utilized pgltools intersect to find loops within 20kb of the eQTL-eGene pair. For each genomic feature, we performed a Fisher's exact test across multiple chromatin loop imbalance p-value thresholds to determine if the genomic feature was enriched in HTALs over the union loop set. To obtain a p-value threshold HTAL set, we filtered all chromatin loops to those exhibiting allelic imbalance with a p-value less than or equal to the threshold.

CNV type analyses

To measure enrichment of CNV types within union loops and HTALs, we identified all CNVs from DeBoever *et al*[42] present in these individuals (1767 deletions and 1045 duplications). We then identified all loops which contained CNVs of the same type using pgltools findloops and intersect1D. Finally, to obtain p-values, we used a binomial approximation to a normal distribution, and tested for an enrichment in

duplications above the genome-wide rate (μ=0.37: the fraction of detected CNVs that were duplications).

Concordance between loop and molecular phenotype imbalance

To examine the relationship between molecular phenotype (RNA-seq and H3K27ac ChIP-seq) allelic imbalance and chromatin loop imbalance, we compared allelic differences in molecular phenotype data to chromatin loop imbalance frequencies in iPSC-CM data. We first removed chromatin loops containing imprinted genes or CNVs. Next, for each union chromatin loop, we utilized the aforementioned allelic imbalance data; for each molecular phenotype, we pooled the iPSC-CM reads from all samples for each individual, applied WASP[81] to reduce reference allele mapping bias, and used MBASED to obtain major allele frequencies of each gene/peak. We then identified the most imbalanced SNV in each gene/peak, and used the SNV's phase to determine the maternal allele frequency of the gene/peak. We then converted maternal allele frequencies to fold changes by dividing the maternal allele frequency by the paternal allele frequency for both molecular phenotypes, and the chromatin loop data.

Chapter 2, in full, is a reprint of the material as it appears in Nature Communications, 2019, William W. Greenwald, He Li, Paola Benaglio, David Jakubosky, Hiroko Matsui, Anthony Schmitt, Siddarth Selvaraj, Matteo D'Antonio, Agnieszka D'Antonio-Chronowska, Erin N. Smith, Kelly A. Frazer. The dissertation author was one of the primary investigators and authors of this paper.

CHAPTER 3 CHROMATIN CO-ACCESSIBILITY IS HIGHLY STRUCTURED, SPANS ENTIRE CHROMOSOMES, AND MEDIATES LONG RANGE REGULATORY GENETIC EFFECTS

Abstract

Chromatin accessibility identifies active regions of the genome, often at transcription factor (TF) binding sites, enhancers, and promoters, and contains regulatory genetic variation. Functionally related accessible sites have been reported to be co-accessible; however, the prevalence and range of co-accessibility is unknown. We perform ATAC-seq in induced pluripotent stem cells from 134 individuals and integrate it with RNA-seq, WGS, and ChIP-seq, providing the first long-range chromosome-length analysis of co-accessibility. We show that co-accessibility is highly connected, with sites having a median of 24 co-accessible partners up to 250Mb away. We also show that co-accessibility can *de novo* identify known and novel co-expressed genes, and co-regulatory TFs and chromatin states. Finally, we perform a *cis* and *trans*-caQTL, a *trans*-eQTL, and examine allelic effects of co-accessibility, identifying tens of thousands of *trans*-caQTLs, and showing that *trans* genetic effects can be propagated through co-accessibility to gene expression thereby affecting cell-type and disease relevant genes.

Introduction

Regulatory genetic variation that affects gene expression and human disease is often found within accessible chromatin sites[41,82-86]. These accessible sites, measured by either DNase-seq[87,88] or ATAC-seq[89,90], identify functional regions of the genome

including active promoters and enhancers[77,88,91], as well as the transcription factor (**TF**) binding sites within them[88,92-94]. However, it is difficult to determine the function of accessible sites as they can be distal from their targets[82]. In order to identify the functionality of accessible sites, previous studies have examined co-accessibility: the coordination of specific chromatin accessibility sites. These studies have examined co-accessibility at fine-scale (ie specific accessible sites) for local *cis* interactions within 10kb (ie co-binding TFs, promoter regulation, and local enhancer regulation)[85,86,95], and long-range *cis* interactions between 10kb and 1.5Mb (i.e. chromatin looping and distal enhancer regulation)[85,95]. They have mainly applied supervised approaches to show that co-accessibility occurs between regulatory regions and their targets, as well as co-binding TFs, and that the majority of *cis* acting, genetically associated co-accessibility occurs at sites <20kb apart[95]. However, due to computational and statistical power, these studies limited their examination of co-accessibility either to fine-scale resolution and local structure, or higher order properties across long-ranges at low resolution[85]. It is thus unclear if co-accessibility extends to *cis* (ie physical co-regulation such as a TF co-binding or looping) and *trans* (ie sequential co-regulation such as a gene network) relationships across long distances (10s-100s of megabases), how many sites across a chromosome are co-accessible with one another (ie how highly connected is co-accessibility), and whether accessible sites can mediate genetic effects on highly distal sites via co-accessibility. A more comprehensive understanding of the co-accessible chromatin landscape and its genetic associations could provide novel insights into the effects of regulatory genetic variation across short and long distances.

As gene regulation involves many distal regulatory components, it is expected that genetic variation could exert *trans* long range regulatory effects. The omnigenic model[96] of gene regulation has recently estimated that 70% of the heritability of gene expression is due to *trans* effects[97]. However, these genetic effects are thought to have effect sizes orders of magnitude smaller than *cis* effects[97], and as they are distal from their targets, they are extremely difficult to identify (creating a power problem due to multiple-testing burden) and delineate from confounded *cis* effects. One possible solution to this statistical power problem could be to leverage chromatin co-accessibility to reduce search space, as gene expression and accessibility of the gene's promoter are known to be correlated[98]. To overcome confounded *cis* effects, it could be possible to use mediator analyses in which one specifically tests for an intermediate effector rather than two independent associations. Additionally, studies examining chromatin accessibility quantitative trait loci (caQTLs) have found moderate overlap between *cis*-caQTLs and *cis*-eQTLs (~30-40%)[83,85]. Thus, it is possible that co-accessibility could be used to tie regulatory elements to their distal co-regulators or gene targets, and then subsequently identify *trans* genetic effects. As this strategy would greatly reduce the number of variant-target pairs tested for *trans* effects (thus reducing multiple testing burden), it may be possible to observe hundreds or thousands of more *trans* effects than previous studies. Identifying co-accessible chromatin regions across entire chromosomes, and the genetic variation associated with these accessible regions, could therefore better elucidate the extent to which genetic variants exert long range *trans* effects, and how these effects may be mediated via co-accessibility.

Here, we perform ATAC-seq in 152 induced pluripotent stem cells (iPSCs) from 134 individuals from iPSCORE[38,99-101], and integrate this data with available WGS and RNA-seq for the same individuals. We call over 1 million accessible chromatin sites and utilize population-level information to identify co-accessible sites by testing for correlation in accessibility between all sites chromosome-wide. We show co-accessibility is highly connected, with sites being co-accessible with an average of 24 other sites, and can span long distances (up to hundreds of megabases). We then use these significant relationships to create co-accessibility networks, and show that neighbors in these networks are enriched for TF co-binding partners, functionally related TFs, spatially colocalized loci (ie loci in a chromatin loop), and co-expressed genes up to 100Mb apart, and can also be used to infer novel TF functionality. Next, we examine the genetic architecture of co-accessibility by measuring allele specific effects (ASE) and performing one of the largest caQTLs studies to date. We show that genetic effects spread through co-accessibility, with highly connected sites being more likely to have a *cis*-caQTL or exhibit ASE; additionally, strong ASE explains 52% of co-accessible weaker ASE. Finally, we leverage these networks to identify more than 92,000 *trans*-caQTLs greater than 1.5Mb from their target, 9 of which are also *trans*-eQTLs for cell type and disease relevant genes. Overall, our data reveals that chromatin co-accessibility is highly connected, spans the length of entire chromosomes, can *de novo* identify co-regulatory TFs, is a mechanism underlying *trans* genetic effects, and can give insight into *trans*-eQTL mechanisms.

Results

Samples, ATAC-seq data generation, and ATAC peak characterization

To measure chromatin co-accessibility, accessible sites were identified from ATAC-seq performed on 152 iPSC lines. These lines were generated from 134 individual from iPSCORE and have previously been shown to be pluripotent and to have high genomic integrity[38] (Figure 3.1A). We obtained a total of 5.5 billion reads, and after QC, filtering, and merging individual samples (see methods), inspected the quality of this data by examining its overlap and consistency with higher order chromatin structure at low-resolution, chromatin states, and H3K27ac peaks. To examine higher order structure (Figure 3.1B), we compared the correlation between ATAC-seq signal in 500kb bins across chromosome 18 to the correlation in Hi-C (from iPSCORE iPSCs[102]), and observed a similar pattern between the two as previously reported[85]. We next used MACS2 to call ATAC-seq peaks (obtaining a total of 1.01 million peaks), and examined the overlap of chromatin states from the iPSC ROADMAP[77] with the peaks. We found peaks to be enriched for active TSS, transcribed regions, enhancers, polycomb-repressed, bivalent TSS, and bivalent enhancers, and depleted for repressed chromatin (heterochromatin and quiescent chromatin, Figure 3.1C). These findings are consistent with properties of accessible chromatin and known specialized use of bivalent and polycomb chromatin in maintaining iPSC pulirpotency[46-49]. Next, we examined the distribution of ATAC-seq reads at H3K27ac peaks from iPSCORE iPSCs and observed an enrichment at the centers of these H3K27ac peaks (Figure 3.1D). Together, these results show that this ATAC-seq data follows known characteristics of accessible chromatin and cell type specific characteristics of iPSCs.

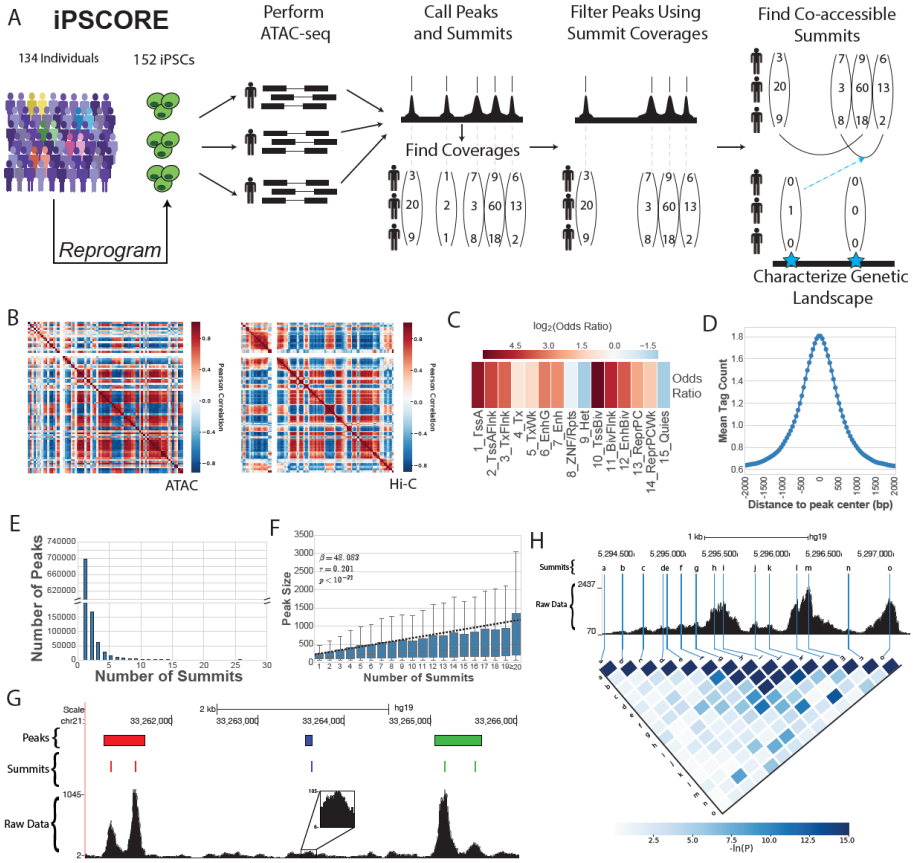

Figure 3.1 Overview and QC of ATAC-seq data

(A) Overview of experimental design. iPSCs from 134 individuals from the iPSCORE cohort were selected for ATAC-seq sequencing (152 samples total). After QC and filtering, all individual's data was utilized to call peaks, and the coverages of each peak and summit were calculated. Peaks were subsequently filtered based on summit coverage, and co-accessible summits were found from summit data. Finally, genetic associations with co-accessibility were examined. (B) Correlation of total reads in 500kb bins for ATAC (left), compared to Hi-C Pearson matrix (right), across chromosome 18. Broadly similar patterns can be observed. (C) Heatmap of the $\log_2$(Odds Ratio) of chromatin state enrichments within ATAC-seq peaks relative to the genome, measured in number of base pairs in each state. (D) Histogram of the average tag count of ATAC-seq reads at H3K27ac peaks across all samples. (E) Histogram showing the number of peaks that have a given number of summits. (F) Boxplot (middle two quantiles and median shown as box and line within box; outliers not shown) and regression line on raw data (dashed line) for the number of summits vs the length of the peak in base pairs. (G) Genome browser picture of the combined ATAC-seq data across all individuals (raw data), summit Calls (Summits), and peak calls (Peaks). Three peaks were called in this window, shown in red, blue, and green; summits are colored by the peak to which they belong. Both the red and green peak calls contain two seemingly distinct peaks, which were identified by their summit calls. The blue peak, while lower, is still peak-shaped and has high coverage (105 reads). (H) Genome browser picture for the combined ATAC-seq data across all individuals (raw data) at a single peak call with 15 summits (labeled a-o). Lines for summits are extended through the raw data, and connect to their label on the heatmap. Heatmap shows the negative natural log of the p-value of correlation between these summits. Correlation quickly decays as a function of distance, and notably, most summits are not strongly correlated.

Peaks contain multiple non-co-accessible summits

As part of ATAC-seq data processing, reads are used to call peaks and their sub-peak structure (**summits**). As summits represent individual TF binding sites within peaks (https://github.com/taoliu/MACS), we examined the number of summits within peaks and found that while the majority of peaks contained a single summit, 31% (315,901) contained multiple summits, with some containing up to 26 (Figure 3.1E). Additionally, we found a strong relationship between the length of peak and the number of summits identified (Pearson Correlation $p < 10^{-32}$; Figure 1F). These patterns are consistent with MACS2's documentation (https://github.com/taoliu/MACS) stating that nearby individual binding sites are called as summits and binned together as a single peak call (Figure 3.1G). We tested whether the summits acted independently by examining the correlation between summit heights in the same peak across individuals, and found that 97.5% of summits were not significantly correlated with other summits within the peak (see methods). Further, the significance of correlation between summits within the same peak decayed with distance (Figure 3.1H). Together, these data indicate that many peak calls contained multiple independent accessible sites; we therefore utilized the 1.21 million summits ATAC summits from peaks that passed QC for downstream analyses of accessible chromatin sites.

Co-accessibility is predominantly distal and highly connected, spanning entire chromosomes

We set out to characterize the local and long-range co-accessibility landscape at fine-scale resolution (ie site-by-site co-accessibility) for each of the 22 autosomes chromosome-wide. We tested the quantile normalized trimmed mean of M values (**TMMs**[52]) of coverage for each site with every other site pairwise on each chromosome using a Linear Mixed Model to account for covariates and kinship. For each pair, we obtained a regression coefficient (**β**) and a p-value. We performed FDR correction of the p-values by chromosome, obtaining between 45 thousand and 3 million significant co-accessible relationships per chromosome (FDR $q < 0.05$). We observed similar numbers of co-accessible pairs normalized to chromosome length, except for chromosome 19 which was ~4x higher. This increase may have been driven by a large cluster of Znf genes known to be highly coordinated[103]. We first examined the distribution of the number of sites each site is co-accessible with (**connectivity**) across all sites for each chromosome, and found the level of connectivity to vary (Figure 3.2A), ranging from sites with no co-accessible partners to those with ~3,000, and a mean of 24.46 partners. Surprisingly, we found that 96.6% of all ATAC sites were co-accessible with at least one other site, suggesting that the vast majority of regulatory sites interact with at least one other regulatory site. We next measured the distances between co-accessible sites (Figure 3.2B). As expected, the most commonly observed distances (i.e. modes of the data) were within the ranges previously studied for local, likely *cis*, co-accessibility (<1.5 Mb Figure 2B). However, the vast majority of co-accessible sites were further than 1.5Mb apart, with some pairs extending up to 250Mb distal from one another and a mean distance of 48.94Mb. Additionally, we found the strength of association to be consistent across distances. Finally, to better understand these data, we visualized the specific site-by-site

correlations across chromosome 18 (Figure 2C), and found that, as Figures 2A and 2B suggested, sites on opposing ends of the chromosome were co-accessible. Further, we zoomed in to eight different resolutions, and at each resolution, we found co-accessible sites spanning almost the entire window (Figure 3.2C). Together, these data indicate that fine-scale co-accessibility extends beyond the local *cis* structure of 1.5Mb that has been previously examined, to sites that are highly distal from one another and likely *trans* in nature (up to hundreds of Mb). Overall, these data reveal that co-accessibility is highly distal and inter-connected.

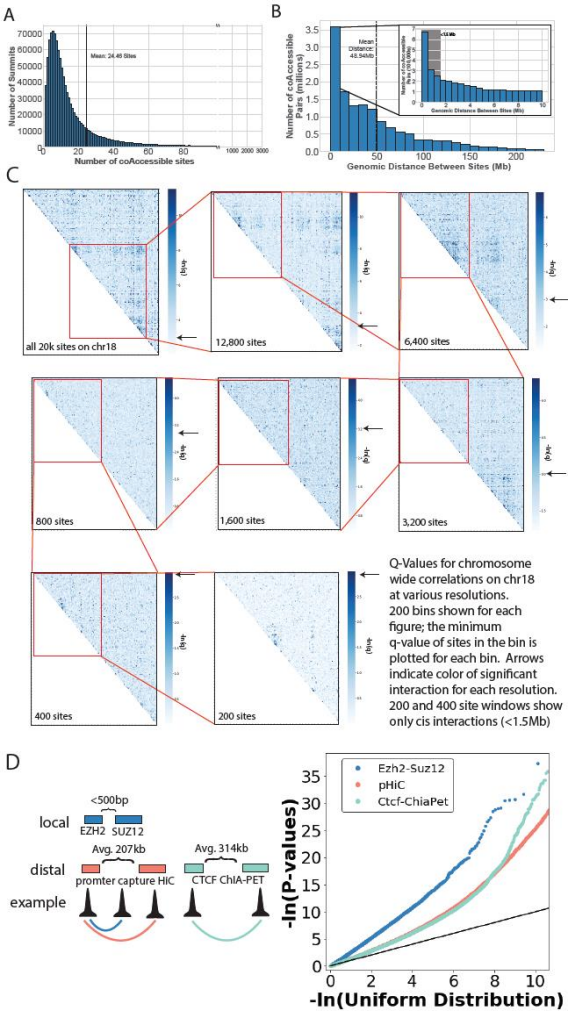

Figure 3.2 Co-accessibility spans entire chromosomes
(A) Histogram showing the number of sites with a given number of co-accessible partners. The mean of 24.46 partners is highlighted by the dashed black line. (B) Histogram showing the distance between sites that are co-accessible. The mean of 48.94Mb is plotted with the dashed black line. The area highlighted in gray shows all co-accessible pairs at distance <1.5 Mb (ie previously studied distances). (C) Heatmap of regression q-values for various resolutions, ranging from the whole chromosome (top left) to 200 sites (bottom right), with 200x200 bins in each heatmap. Each box in red shows the region on the next zoom, starting in the top left and snaking to the bottom right. Each resolution is double the number of sites from the corresponding finer resolution. In the bottom right, each pixel in the heatmap is a single site (the grid is 200x200); in other panels, each pixel is the most significant q-value for all sites within the bin (ie for 400 sites, each pixel is 2-sites; for the entire chromosome, each pixel is ~70 sites). Arrows on each color bar indicate color of a significant correlation (q = 0.05). At each resolution both local and distal significant associations can be seen. (D) QQ plot showing enrichments of p-values of co-accessibility combined across all chromosomes. (Left) Cartoon showing types of sites within each class: Ezh2-Suz12 TF pairs within 500bp of one another (blue), promoter capture HiC (red; avg 207kb apart) and CTCF ChIA-PET (teal; avg 314kb apart). Enrichments for local *cis* (blue) are above those for long-range *cis* (teal and red).

Co-accessible sites are enriched for known biological processes

To determine whether calling co-accessibility across entire chromosomes reduced our power to identify *cis* co-accessibility, and/or resulted in technical artifacts, we examined the enrichment of co-accessibility for previously associated *cis* biological processes. First, using all co-accessible pairs from all chromosomes, we measured the enrichment for a local *cis* process: TF co-binding partners in the same protein complex (*EZH2* and *SUZ12* from PRC2[104,105]). We found that *EZH2-SUZ12* sites within 500bp of one another were highly enriched (Figure 3.2D, blue). Next, we examined enrichment for a long-range *cis* process: chromatin looping[86]. We found that opposing anchors of promoter centric chromatin loops measured via iPSC promoter capture HiC[76] (mean 207kb apart; Figure 2D red), as well as structural chromatin loops measured via CTCF ChIA-PET[16] from GM12878 (mean 314kb apart, Figure 2D teal), were also enriched. However, the enrichment for the long-range looping was lower than the local protein co-binding, consistent with *cis* effects being distance-dependent. Overall, these results show that the subset of co-accessibility within previously studied distances (<1.5Mb) recapitulates known *cis* characteristics of iPSCs.

Co-accessibility intrinsically captures and predicts co-expression and chromatin state interactions

We performed unsupervised analyses to examine whether co-accessibility captures co-regulatory regions and genes, and if these relationships could be learned directly from co-accessibility. Since gene/protein regulation naturally has a network-like structure[106], we modeled the co-accessibility as a network (Figure 3.3A). For each

chromosome, we built an undirected graph with accessible sites as nodes, and edges between FDR q < 0.05 co-accessible sites weighted by their regression coefficient. We then annotated each site based on its overlap with gene promoters from GENCODE[107], ROADMAP[77] chromatin states for iPSC, and TF binding sites from ChIP-seq data in ESCs[104] (see methods; Figure 3A). We first examined the correlation in gene expression (**co-expression**) for genes with co-accessible promoters using RNA seq data from the 154 iPSC lines[42]. We compared both the enrichment of co-expression correlation p-value (Figure 3.3B), as well as the proportion of tests that were significant (Figure 3.3C), across gene pairs that were stratified by distance up to 100Mb. As expected, we observed a high proportion (35%) of the examined *cis* gene pairs (<1.5Mb) whose promoters were co-accessible to be significantly co-expressed. Interestingly, we also observed a strong-enrichment for co-expression in pairs of genes with co-accessible promoters that were highly distal to one another (at least 10-100Mb apart, 27-30%). This enrichment was far greater than random distance matched genes (3%, Figure 3.3C, dashed line). Further, we found the proportion of significant tests to decay more quickly for genes within 1.5Mb of one another (slope = -0.26), compared to genes between 1.5Mb and 10Mb apart (slope = -0.03), suggesting the observed co-expression was largely driven by *trans* effects (as *cis*, but not *trans*, effects would be expected to decay with distance. These results reveal that co-accessible sites can *de novo* identify co-expressed gene pairs, and that while co-expression occurs most frequently in *cis*, it also occurs frequently across long ranges (including hundreds of megabases).

We next used the chromatin state annotations (Figure 3.3A) to examine co-accessibility between states (see methods). Due to computational constraints we focused on chromosome 18, observing three distinct clusters which highlighted known chromatin state interactions and iPSC specific biology[77,46,47,49,102] (Figure 3.3D): 1) genic enhancers and transcribed chromatin (active or weak); 2) enhancers, bivalent enhancers, and TSS flanking chromatin ; and 3) active TSSes, bivalent TSSes, repressed/weak repressed, and heterochromatin . In addition to these 3 clusters, we found crossover between: 1) two different clusters (clusters 2 and 3) through Promoter-Promoter-Flanking interactions; and 2) two different subclusters (repressed and promoter in cluster 3) through active and bivalent TSS interactions with either strong or weak repressed polycomb . Overall, the observed gene co-expression and chromatin state clustering from unsupervised analyses suggest that co-accessibility can be used to *de novo* identify co-regulatory genes and chromatin states.

Co-accessibility identifies novel co-regulatory TFs, as well as distance-dependent TF co-regulation

We sought to identify novel TF co-regulatory information captured by co-accessibility, and use it to derive new insights into the transcription factor landscape of iPSC gene regulation. Using the 51 ChIP-seq TF annotations on the co-accessibility networks (Figure 3.3A), we examined which pairs of TFs tended to be co-accessible more often than by chance (see methods). These transcription factor pair enrichments separated into five main clusters, (Figure 3.3E) each of which contained numerous TFs that were known to be co-regulatory or functionally related: 1) pluripotency factors,

including OCT4 (POU5F1), NANOG, and TEAD4 ; 2) cell proliferation and organogenesis related TFs, including BRCA1, JARID1A, FOSL1, and SIX5 ; 3) transcription and proliferation, including CtBP2, GABP, SP4, CHD2, and SRF for transcription, and c-Myc, AFT3, MXI1 and NRF1 for proliferation ; 4) chromatin loop factors/structural factors, including Rad21, CTCF, YY1, SP1, JUND1, and Znf143 ; and 5) transcription, including the promoter binding factors Pol2, TAF1&7, RBBP5, and TBP . While these five clusters recapitulated known TF groupings and functionality, we identified novel functions for TFs from cluster membership, subclustering, and cluster cross-over. As an example for cluster membership, RFX5 was a member of the proliferation/growth cluster, suggesting it may play a role in cancer; this is consistent with previous studies that found RFX5 upregulated in liver cancer, which results in the activation of genes associated with poor prognosis[108]. As a subclustering example, although Znf143 has been observed in promoter enhancer loops, it was not a member of the subcluster of the promoter enhancer specific loop TFs JunD, YY1, and SP1[109,110]; rather, it had patterns similar to the broad loop factors Rad21 and CTCF, suggesting it may play a broad role in loop formation. As a crossover example, in the pluripotency cluster, we found some TFs (ex TEAD4, HDAC2) to have cross-over with the loop cluster, and others to not cross-over (ex BCL11A, NANOG), suggesting that some factors may have a more distal regulatory role than others. Overall, these analyses reveal that chromatin co-accessibility *de novo* recapitulates known gene regulation patterns and interacting TFs (including those that are cell type specific), and suggests that it may be possible to infer TF functionality from co-accessibility data.

As the examined networks contained co-accessible pairs at many different distances (from kilobases to megabases), we sought to find TF interactions unique to different regulatory distances. We stratified the network to interactions within 10kb (**local *cis***; Figure 3F), between 10kb and 1.5Mb (**long-range *cis***; Figure 3G), and greater than 1.5Mb (**distal**; Figure 3H). For the local and long-range *cis* networks, we used all chromosomes; for distal, we utilized chromosome 18 due to computational constraints. Across all three distance-stratified networks, we observed promoter binding, proliferation, pluripotency, and looping clusters; however, the specificity of these clusters, and the cross-overs between them, were different. For instance, in the local *cis* network, looping only contained CTCF and Rad21, whereas in the long-range *cis* network, the looping cluster also included Znf143, suggesting that Znf143 may act only on one anchor side for loop formation. Further, in both *cis* networks, the promoter cluster was separate from the loop cluster, consistent with only some of *cis* regulation involving chromatin looping; however, in the distal network, the promoter and loop clusters were combined, consistent with the majority of highly distal *cis* regulation utilizing chromatin loops. This suggests that the distal network contained *cis* interactions in addition to the interactions which span hundreds of megabases and are likely *trans*. We also observed stronger cross-over between the proliferation and promoter clusters in both the long-range *cis* and distal networks, compared to the local *cis* network, suggesting these TFs (ex JARID1A, GTF2F1) may mainly act through long-range and/or *trans* regulation. Finally, we observed more negative associations in the distal network than either *cis* network, suggesting that *trans* regulation may be comprised of more antagonistic

regulation than *cis* regulation. Overall, these analyses suggest that TFs have different co-regulatory partners and directional relationships across different distances.

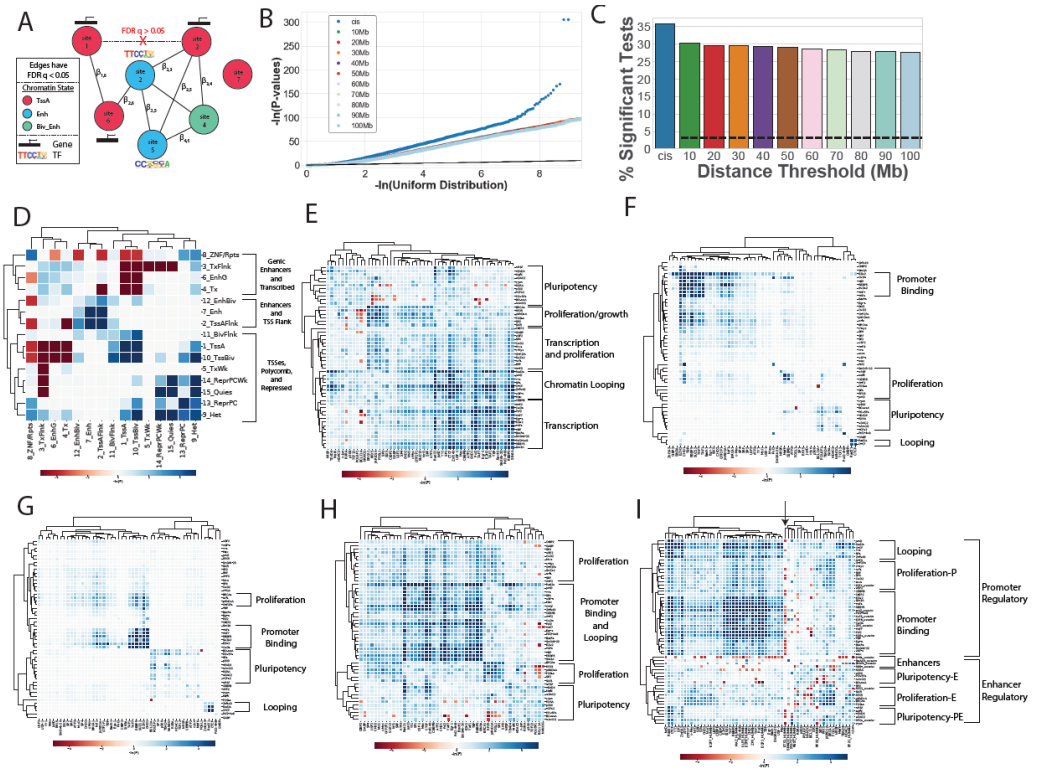

Figure 3.3 Modeling co-accessibility as a network
(A) Cartoon of a co-accessibility network. Accessible sites are represented as nodes, and FDR q < 0.05 co-accessible relationships are edges. For example, site 1 and 6 have a significant output from the LMM and thus have an edge; 1 and 3 do not. Edges are weighted by their regression coefficient (β). Nodes are labelled by their chromatin state (shown as colors), the genes whose TSS they overlap (shown as black boxes with arrows), and by the TF ChIP-seq peaks from ESC they overlap (shown as motifs). (B) QQ plot showing enrichments of p-values of correlation for co-expression of genes whose promoters are neighbors in the co-accessible network. Gene pairs are stratified by those that are at most 1.5Mb apart (*cis*) or those that are at least a given distance apart (colors; overlapping stratifications). *Cis* is enriched above all other distances, which are overlapping. (C) Bar plot showing the percent of co-expressed pairs that are FDR q < 0.05 at each distance threshold. Colors are shared between (B) and (C). (D-I) Heatmaps showing the signed empirical p-values of connectivity for (D) chromatin states, (E-H) TF ChIP-seqs, or (I) TF ChIP-seqs and predicted motifs. All distances (*cis* through 100Mb) shown for D, E, and I. (F) uses the local *cis* subnetwork induced from sites that are within 10kb of one another. (G) uses the long-range *cis* subnetwork induced from sites between 10kb and 1.5Mb apart. (H) uses the *distal* subnetwork induced from sites that are at least 1.5Mb apart. (D), (E), and (I) use chromosome 18. (F) and (G) use all edges from the genome-wide network. Clusters are labelled using the most common functionality of the included genes.

Incorporation of motif predictions in identifying TF co-regulation reveals distinct promoter and enhancer clusters

To further examine what novel TF biology could be learned from co-accessibility, we expanded the network TF annotations to include predicted binding sites for TFs that were expressed and enriched in iPSCs (Figure 3.3I). We identified two superclusters which were composed of seven clusters that we named as follows: looping, promoter centric proliferation (proliferation-P), promoter binding, enhancers, promoter and enhancer centric pluripotency (pluripotency-PE), proliferation, and enhancer centric pluripotency (pluripotency-E). The two superclusters were separated by ETS1, with the top supercluster (containing Pol2) having a negative association. ETS1 has been shown to be a TF at sites occupied by Pol3 and involved in enhancer RNA transcription[111]. These observations suggest that TFs in the three clusters composing the supercluster anti-associated with ETS1 are primarily located at promoters, and TFs in the other supercluster (composed of four clusters) are primarily located at enhancers. Interestingly, we found proliferation clusters both in the promoter (ex TFs: NRF1, SRF) and the enhancer (ex TFs: JARID1, BRCA1) superclusters, as well as two different pluripotency clusters within the enhancer supercluster (ex TFs: OCT4, NANOG in pluripotency-E; TEAD4, HDAC2 in pluripotency-PE). This suggests that TFs with similar functions may act across different genomic distances. While most of the promoter supercluster TFs were anti-associated with ETS1, and most of the enhancer supercluster TFs were not associated with ETS1, one cluster in the enhancer supercluster was anti-associated with ETS1 (Pluripotency-PE). This cluster (pluripotency-PE) also had a strong cross-over with the loop cluster, which in turn has a cross-over with the promoter binding cluster,

suggesting that the TFs in the pluripotency-PE cluster regulate locally at promoters and distally through looping. Overall, these results show that co-accessibility can help delineate TFs that have primarily distal regulatory roles (Pluripotency-E and Proliferation-E clusters) from those that have primarily promoter regulatory roles (Proliferation-P and Promoter Binding clusters) from those which do both (Looping and Pluripotency-PE clusters).

Identification of caQTLs and relation to co-accessibility

We sought to provide an in-depth characterization of how genetics is associated with total accessibility of sites (QTLs), as well as allele specific effects (ASE), in the context of *cis* and *trans* effects. We obtained genotypes for the 134 individuals from iPSCORE that had been previously identified using 50X WGS[42], and tested for associations between the height of the accessible site and all genetic variants within 100kb[83,85] of it using a linear mixed model. Across all chromosomes, we found 235k sites with an associated genetic variant within 100kb (***cis*-caSites**) with an FDR $q < 0.05$ (21%), which is consistent with previous estimates of the fraction of accessibility that is explained by variation[85]. We examined the enriched motifs at these *cis*-caSites, and found the top motifs enriched to be OCT4, CTCF, NANOG, SOX-family, and TEAD-family, consistent with iPSC gene regulation . We next examined the chromatin states enriched at these sites, and found an enrichment for non-promoter chromatin states (Figure 3.4A), suggesting that genetically associated sites were more likely to be distal regulatory in nature than located at gene TSSes or flanking chromatin. Next, we examined the association between co-accessibility and genetics by measuring the proportion of sites

with a given connectivity that were significant *cis*-caSites (Figure 3.4B). We found that higher site connectivity corresponded to a higher proportion of significant *cis*-caSites, indicating that having more co-accessible partners increases the likelihood of having a *cis* genetic variant. This result suggests that having co-accessible partners allows for compensation against changes in a site due to genetic effects. Overall, these analyses identify sites whose total accessibility is genetically associated, and show that they are more likely to occur at iPSC distal-regulatory elements and to have high connectivity.

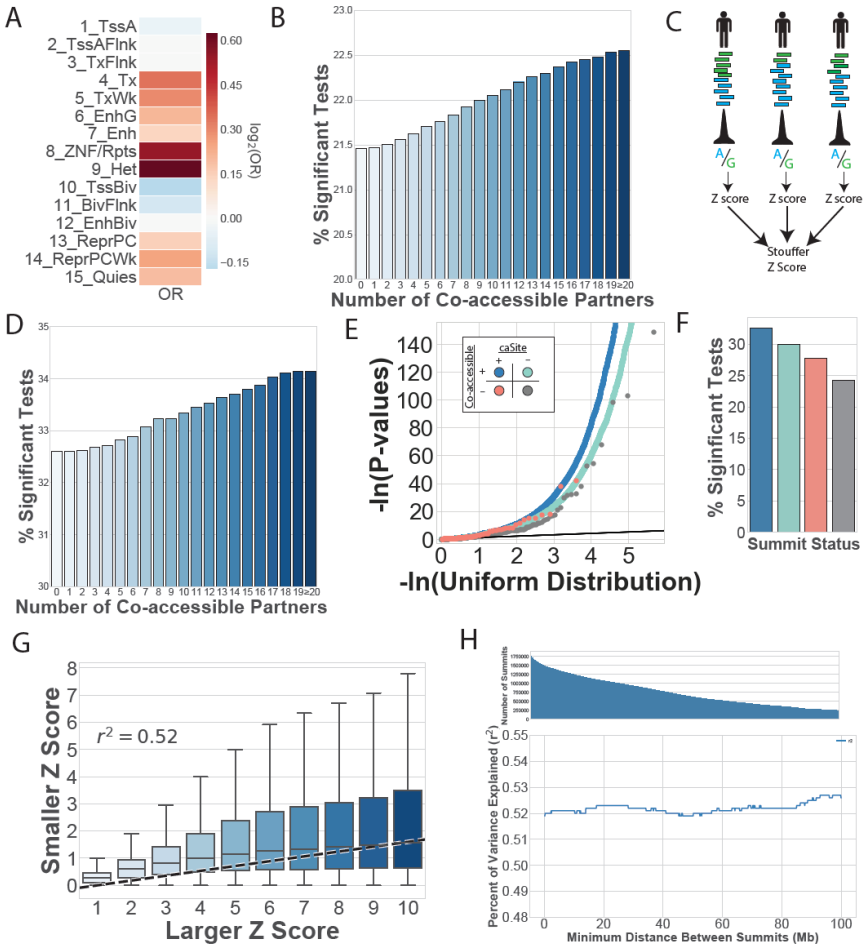

Figure 3.4 Co-accessibility and genetic associations

(A) Heatmap showing the $\log_2$ of the odds ratio of enrichment for chromatin states at *cis*-caSites compared to all accessible sites, with odds ratios set to 1 if the enrichment was non-significant. (B) Barplot showing the proportion of significant caSites as a function of the number of co-accessible partners in their networks. As sites have more co-accessible partners, they are more likely to have a *cis*-caQTL. Bars are colored by connectivity. (C) Workflow for identifying ASE for chromatin accessibility. Individual imbalance measurements were obtained per individual at sites with at least 10 reads. Z scores were calculated, and then combined across individuals for a single meta Z score per site. (D) Barplot showing the proportion of significant sites with ASE as a function of connectivity. As sites have more co-accessible partners, they are more likely to have exhibit ASE. Bars are colored by connectivity. (E) QQ-plot and (F) Proportion of significant tests for ASE at sites that are co-accessible and caSites (blue), co-accessible and non-caSites (teal), singleton and caSites (red), or singleton and non-caSites (grey). Differences associated with caSites status can be seen by comparing blue to teal and red to grey. Differences associated with co-accessibility can be seen by comparing blue to red and teal to grey. (G) Boxplot of Z score relationship between co-accessible sites where both sites have heterozygous variants. Larger Z scores are plotted on the x-axis, and the paired smaller Z-score is on the y-axis. The regression line (calculated on the raw data) is plotted as a dashed black line. (H) Line plot showing the r^2 value of larger ASE predicting co-accessible smaller ASE (as in panel G), restricting the analyses to sites at least X distance apart (up to 100Mb). The r^2 holds consistent across all distances.

Co-accessibility explains a large fraction of variation in ASE

To further probe the extent to which genetic effects were mediated by co-accessibility, we examined ASE. We measured ASE by calculating imbalance within each individual at all heterozygous variants within 200bp of an accessible site, and then meta-analyzing across individuals using Stouffer's method (Figure 3.4C). This analysis identified >48,000 significant ASE sites at an FDR $q < 0.05$ – notably, while ASE identifies regions associated with genetic effects, it does not delineate whether the imbalanced variant is causal for the effect, or neutral but in phase with a causal variant (ie proxy variant). We examined if sites with higher connectivity were more likely to exhibit ASE, and found that the proportion of significant ASE sites increased with connectivity (Figure 3.4D). This result suggests that co-accessibility may allow for a *cis* genetic effect to be mediated in *trans* to a co-accessible site (ie co-accessibility could be one of the processes which causes a proxy variant to be imbalanced).

To compare the relative effects of co-accessibility and *cis* genetic variation on ASE, we next we compared the distribution of ASE p-values (Figure 3.4E) and proportion of significant tests (Figure 3.4F) across four distinct sets of sites: 1) co-accessible *cis*-caSites (blue); 2) single *cis*-caSites (red); 3) co-accessible non-caSites (teal); and 4) single non-caSites (grey). As expected, we found both sets of caSites to be more enriched for ASE than their non-associated counterparts (Figure 3.4E,F blue vs teal, and red vs grey). Additionally, we found both sets of co-accessible sites to be more enriched for ASE than non-co-accessible sites (Figure 3.4E,F, blue vs red, and teal vs

grey), consistent with Figure 4D. Interestingly, we found co-accessibility status to be more enriched for ASE than *cis*-caSite status (Figure 3.4F, both blue and teal are enriched above red). This result further supports *trans* genetic effects being mediated through co-accessibility, as these sites do not have a significant *cis*-caQTL, but do show significant allelic effects. To make sure that this observation was not predominantly driven by false negative caQTLs, we examined whether ASE in one site could explain ASE in co-accessible sites by regressing the lead Z score of a co-accessible network against each partner Z score en masse across all chromosomes simultaneously (see methods). We found a large fraction of variation in ASE to be explained by the single most imbalanced co-accessible site (r^2=0.52, $p < 10^{-32}$; Figure 4G), showing that genetic variants can exert *trans* effects via co-accessibility. Finally, we examined whether this high predictability was consistent across varying genomic distances. We found the r^2 to hold surprisingly constant up to 100Mb apart, ranging between 0.51 and 0.53 (Figure 3.4H), despite a large difference in the number of pairs used in the model (between 250k and 1.25M). Together, these results suggest that an accessible site can be influenced in *trans* by a distal genetic variant through intermediate effects on a co-accessible partner.

Identification of trans-caQTLs by leveraging co-accessibility network

We set out to identify genetic variants that indirectly affect distal sites by mediating their *cis* effects through co-accessibility (ie *trans*-caQTLs). To identify *trans*-caQTLs within the same chromosome, we leveraged the co-accessibility network to perform targeted association tests, thereby reducing multiple hypothesis testing. To perform these analyses, we restricted our tests to variants that were *cis*-caQTLs, and

tested them against neighbors of the respective *cis*-caSite in the co-accessibility network that were at least 1.5Mb away (Figure 3.5A). This identified 368,639 putative *trans*-caQTL-caSite pairs out of a tested 9,967,402 pairs (3.7%) at an FDR $q < 0.05$. Notably, many of these putative *trans*-caQTLs were highly distal to their targets (Figure 3.5B), with some hundreds of megabases away. However, this regression analysis cannot delineate a true *trans* interaction in which the variant's effect on an accessible site is mediated through a co-accessible partner (Figure 3.5C left) from two independent *cis* effects driven by the same variant (Figure 3.5C right). We thus further probed these putative *trans*-caQTLs by performing a mediator analysis to identify variants with a statistically significant fraction of their association with the *trans*-caSite explained by the height of the *cis*-caSite[112]. For these analyses, we tested all genotypes in the 134 individuals (ie if the site was multiallelic, we included all alleles; 80% of multiallelic sites were indels). Out of the 934,136 putative *trans*-caQTL genotypes, the mediator analysis found 92,638 to be significantly mediated at an FDR $q < 0.05$ (9.9% of the putative, ~1% of all tests)As sites have high connectivity, it is possible that a given cis-caSite which mediates a *trans* effect (*cis*-mediator-caSite) exerts its effects on multiple co-accessible partners. We thus examined whether *cis*-mediator-caSites affected multiple co-accessible *trans*-caSites, and found the 92,638 *trans*-caSites to be mediated by 29,362 *cis*-mediator-caSites, with a mean of 3.16 *trans*-caSites per mediator (Figure 3.5D). This suggests that when a variant affects an accessible site, the effects are mediated throughout its co-accessibility network, rather than in a pair-wise fashion with only one of the co-accessible partners. We compared the motifs underlying *cis*-mediator-caSites and *trans*-caSites, and found both to be similarly enriched for OCT4, CTCF, NANOG, SOX-

family, and TEAD-family motifs. We next examined the chromatin states at *cis*-mediator-caSites and *trans*-caSites, and found both to be enriched for promoter and gene centric chromatin states (Fisher's Exact FDR $q < 0.05$; Figure 5E); however, only *trans*-caSites were depleted at enhancers and bivalent enhancers. The fact that these enrichments are different suggests that genetic effects have a directionality within co-accessibility networks. Together, these analyses identified tens of thousands of *trans*-caQTLs, suggests that *trans* effects are directionally mediated throughout a network, and shows that co-accessibility can be leveraged to identify *trans*-caQTLs from a relatively small sample size.

To gain better insight into the mechanisms underlying these *trans*-caQTLs, we characterized two large co-accessibility networks centered on *cis*-mediator-caSites. These networks were chosen because their *cis*-mediator-caSite was at the promoter of a TF, and their *trans*-caSite contained a binding site for that TF. The smaller of these two networks was on chromosome 17 (Figure 3.5F) with 77 total sites, 30 of which overlapped gene promoters. One of these 30 sites was at the RARA gene promoter, and was a *cis*-mediator-caSite whose *trans*-caSite overlapped a RARA binding motif. Four other sites were also at RARA binding sites. The RARA gene has been implicated in development, differentiation, and transcription of clock genes[113]; we therefore examined the network for iPSC TFs, and found three sites overlapping ChIP-seq binding sites for the core pluripotency TFs (OCT4, NANOG, and TEAD4). Finally, we examined the function of the 30 genes whose promoters were in the RARA network, and found their proteins to be statistically enriched for having protein-protein interactions (**PPIs**) in StringDB

(StringDB enrichment $p = 9.65x10^{-3}$). Further, the functionality of these genes was enriched for gene sets for multiple cancer types, including acute myeloid leukemia (AML) and Breast Cancer, suggesting co-accessibility network dysregulation may play role in cell-type relevant disease. The second example, on chromosome 18, was a co-accessibility network comprised of 261 sites, of which 130 were at gene promoters (Figure 3.5G). One of these 130 sites was at the PLAG1 promoter, and was a *cis*-mediator-caSite for a *trans*-caSite at a PLAG1 motif. In addition to the *trans*-caSite, 23 other sites contained PLAG1 motifs. PLAG1 is developmentally regulated[113]; we therefore also examined iPSC TFs, and found 24 sites overlapping a ChIP-seq peak for NANOG, OCT4, or TEAD4. Finally, the proteins transcribed by the genes in the network were significantly enriched for being having PPIs (StringDB enrichment $p = 8.12x10^{-4}$), but not for any StringDB gene sets. Notably, this set proteins contained 19 experimentally validated PPIs. Overall, these analyses show that co-accessibility can be used to identify novel *trans* regulatory modules which can be disease-associated.

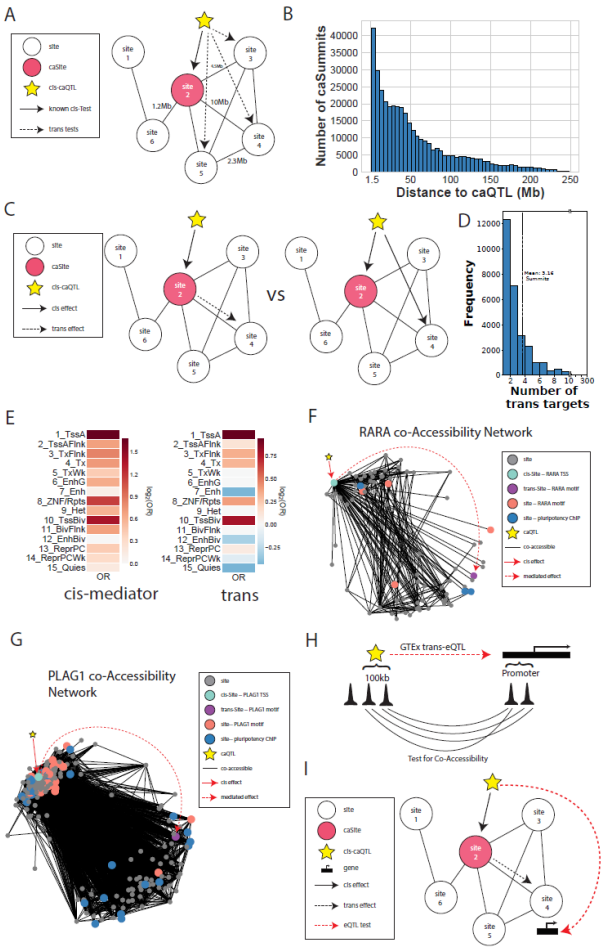

Figure 3.5 *trans* ca-QTL and e-QTL

(A) Cartoon illustrating how putative *trans*-caQTLs were tested. *Cis*-caQTLs (star) were tested as *trans*-caQTLs against the neighbors (sites 3,4, and 5) of the *cis*-caSite (site 2) in the co-accessibility network. (B) Histogram showing the distance distribution of significant putative *trans*-caQTLs to their target *trans*-caSite. (C) Cartoon illustrating mediator analysis, which tests whether a variant exerts a *trans* effect on a site (site 4) through an intermediate site (site 2; left), or whether the variant exerts two independent effects (right). (D) Distribution of the number of mediated *trans*-caSites per *cis*-mediator-caSite. (E) Heatmaps showing the $\log_2$ of the odds ratio of enrichment for chromatin states at (left) *cis*-mediator-caSites or (right) *trans*-caSites compared to all *cis*-caSites, with odds ratios set to 1 if the enrichment was non-significant. (F and G) Co-accessibility networks centered on two particular sites: (F) an accessible site at the RARA promoter, and (G) an accessible site at the PLAG1 promoter. Nodes are accessible sites, edges show significant co-accessibility. Node color indicates whether the node is the *cis*-medaitor-caSites which defines the network (teal), the *trans*-caSite containing the TF for the respective gene (purple), a non-associated site containing the respective TF (red), a site containing an iPSC ChIP-seq factor (blue), or a different site (grey). The caQTL is shown as the star, with a solid red line for its *cis*-effect, and a dashed red line showing the mediated effect from the *cis*-mediator-caSite to the *trans*-caSite. (H) Cartoon illustrating tests for co-accessibility at GTEx *trans*-eQTLs. Sites within 100kb of the *trans*-eQTL variant were tested against sites at the eGene promoter. (I) Cartoon illustrating how *trans* eQTLs were tested. Sites with a significant mediator q value (site 2) had their *cis*-caQTLs (star) tested as an eQTL against the genes which had co-accessible promoters (site 4).

Identification of *trans*-eQTLs from *trans*-caQTLs at promoters

It is possible that some of variants affecting chromatin accessibility could propagate their effects to changes in gene expression. We hypothesized that some of the variants underlying *trans*-caQTLs whose *trans*-caSite was at a promoter for a gene were also *trans*-eQTLs for that gene. As previous studies have shown that a large cohort is required for sufficient power to detect *trans*-eQTLs[84], we initially examined whether *trans*-eQTLs previously identified in GTEx (in different tissues and inter-chromosomally) exhibited co-accessibility with their eGene in our data (Figure 3.5H). We performed a targeted inter-chromosomal analysis (as our networks were all intra-chromosomal), examining the 32 non-MHC *trans*-eQTLs in GTEx for genes expressed in the iPSCORE iPSCs. Surprisingly, we found that 97% (31/32) of these *trans*-eQTLs had co-accessibility between a site at the promoter of the eGene and one near the eQTL (FDR $q< 0.05$, see methods), despite none of these *trans*-eQTLs being discovered in stem cells. This result suggests the majority of *trans*-eQTLs have co-accessibility associated with them, and that co-accessibility may be conserved across cell types.

Next, we sought to determine whether *trans*-caQTLs could inform and increase power for detecting *trans*-eQTLs. We performed an intra-chromosomal *trans*-eQTL by identifying all genes whose promoters overlapped a *trans*-caSite, and used an LMM to test for association between gene expression and the genotype of the *trans*-caQTL (Figure 3.5I). Overall, we found an enrichment within the *trans*-eQTL p-value distribution (λ_{gc}=1.49), and 9 significant *trans*-eQTLs (FDR $q < 0.05$; Table 1). These

results show that co-accessibility can be utilized to increase power in detection of *trans*-eQTLs, as GTEx (a larger multi-tissue study) identified an average of 2.7 *trans*-eQTLs per tissue. The 9 *trans*-eGenes were *RASSF7*, *EHMT1*, *DPP9*, *LMNB2*, *RGS3*, *AC009133.17*, *SDCCAG8*, *PDE2A*, and *NR1D1*, which are all related to iPSC functionality (ie cell cycle, growth, division) or relevant diseases (ie cancer)[113]. The median distance between the corresponding eQTL and eGene was 27Mb; one pair was over 200Mb apart. Within the average 27Mb window of a gene, the 134 individuals in this study had ~400k variants; thus, our approach of only testing *trans*-caQTLs against expression levels of genes with *trans*-caSites at their promoters greatly reduced the p-value threshold for significance. These results demonstrate the advantages of using co-accessibility to identify *trans* effects on cell type specific gene regulation.

Table 3.1 *trans*-eQTL results

Variant	Summit	Gene ID	Gene Name	Beta	P	Q	Variant Summit Distance
chr11:85425376	chr11:560507	ENSG00000099849.10	RASSF7	-3.517668244	8.09E-16	8.29E-12	8486486
chr1:18566851	chr1:243418251	ENSG00000054282.11	SDCCAG8	0.219172129	2.14E-07	0.001098407	22485140
chr16:2553486	chr16:29754764	ENSG00000260719.1	AC009133.17	-0.74599912	4.05E-06	0.013810673	2720127 8
chr9:139667161	chr9:140683708	ENSG00000181090.13	EHMT1	1.774032818	2.07E-05	0.049666034	1016547
chr19:9999210	chr19:4685575	ENSG00000142002.12	DPP9	0.476342794	3.43E-05	0.049666034	5313635
chr19:13023591	chr19:2458281	ENSG00000176619.6	LMNB2	0.51023325	3.86E-05	0.049666034	1056531 0
chr9:80914060	chr9:116263136	ENSG00000138835.18	RGS3	-0.637221951	4.09E-05	0.049666034	3534907 6
chr11:66141427	chr11:72370422	ENSG00000186642.11	PDE2A	-0.974316619	4.17E-05	0.049666034	6228995
chr17:65421333	chr17:38258696	ENSG00000126368.5	NR1D1	0.909156673	4.36E-05	0.049666034	2716263 7

Integrating co-accessible annotations to infer trans-eQTL mechanisms

Finally, to characterize how chromatin accessibility, gene expression, regulatory variation, chromatin states, TFs, connectivity, and genomic distance fit within the context of co-accessibility, we visualized one of the networks from the *trans*-eQTL analysis (Figure 3.6). The network centered on the *cis*-mediator-caSummit (chr17:65456616) contains 12 co-accessible sites, spanning both the P and Q arms of chr17 (Figure 3.6A). The *cis*-caQTL for the *cis*-mediator-caSummit (chr17:65456616; Figure 6C) is mediated to two *trans*-caSites, chr17:38258124 and chr17:75463608 (Figure 3.6B, dashed red lines; Figure 6D). One of these *trans*-caSites, chr17:38258124, is at the *NR1D1* promoter (Figure 3.6B) which is associated with circadian rhythm and reported to have iPSC specific functionality[114]. The *cis*-caQTL is also a *trans*-eQTL for *NR1D1* (Figure 3.6E). In this network, the 11 sites co-accessible with *cis*-mediator-caSummit chr17:65456616 were not co-accessible with one another (Figure 3.6B; the hub and spoke shape of the network). These 11 sites are at 6 different types of chromatin states, 6 gene promoters, and contain numerous TF motifs (Figure 3.6B); and the *cis*-mediator-caSite is at an iPSC weak enhancer (EnhW2 from the 25-state model of E020 in ROADMAP) and contains motifs for PITX2A, RARB, THA, THB, and ZN770. Together, these data suggest that the *trans*-eQTL, which is 27Mb distal from its eGene *NR1D1*, exerts its effects by modulating the binding of one or more of the 5 TFs at the *cis*-mediator-caSite. Overall, these data exemplify how annotating co-accessibility networks with multiple types of molecular phenotypes can identify *trans* genetic effects and putative mechanisms underlying them.

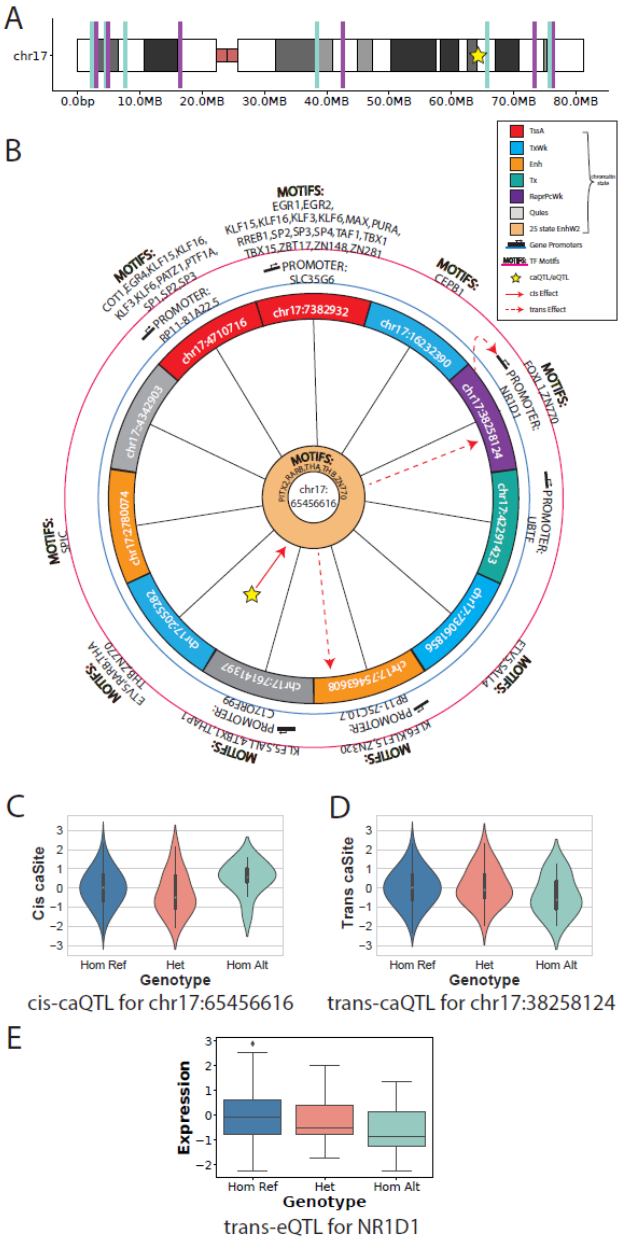

Figure 3.6 The chr17:65456616 co-accessibility network
An overview of the chr17:65456616 co-accessible network. (Top) ideogram showing location of sites across the chromosome. Two colors are used so that nearby sites are visible. The star indicates the caQTL/eQTL. (Middle) The co-accessibility network; nodes shown as the inner circle (chr17:6546616) or partitions of the outer circle (neighbors). Black lines (spokes) are co-accessible relationships, solid red line shows *cis*-QTL effect, and dashed red lines show mediated *trans*-caQTL effects. Colors on the partition show the chromatin state, the blue circle lists gene promoters overlapped, and the red outer circle lists Motifs that are overlapped. (Bottom) Violin or boxplots of QTL results for the *cis*-medaitor-caQTL for crh17:65456616, the *trans*-caQTL for chr17:38258124, and the *trans*-eQTL for NR1D1. All effects are from the same variant.

Discussion

Here, we performed ATAC-seq in 152 iPSC lines from 134 individuals, and use the data to find chromosome-wide co-accessibility. We show co-accessibility is highly connected, with sites being co-accessible with an average of 24 other sites, and can span long distances (up to hundreds of megabases). We then show that, from annotated co-accessibility alone, it is possible to find, *de novo*, co-regulatory chromatin states, genes, and TFs. Additionally, we use this information to infer novel TF functionality and observe that binding sites for TFs with similar functions or that act in complexes are co-accessible at long distances (up to hundreds of megabases). Finally, we perform one of the largest chromatin accessibility QTLS (caQTLs) to date, identifying hundreds of thousands of *cis*-caQTLs, tens of thousands of *trans*-caQTLs, and 9 *trans-eQTLs* as well as putative mechanisms underlying them.

We show that chromatin co-accessibility is a mechanism by which distal *trans* genetic effects are mediated. We found that co-accessible sites were more likely to have a *cis* genetic effect, and that allelic effects were predictive of co-accessible allelic effects. Additionally, we also show that genetic variants that are associated with accessibility often mediate their effects to multiple distal partners through co-accessibility. Together, these results suggest that co-accessibility may function as an insulator, allowing a given regulatory system to be more robust to perturbation by having sites compensate for their co-accessible partners. Future studies examining the effects of perturbing multiple aspects of the same co-accessibility network in a dose dependent manner could validate

this hypothesis, as well as provide insight into the spreading of *cis* genetic effects to *trans* throughout the genome.

Previous studies[84] utilizing large cohorts of individuals and multiple tissue types have shown that it is difficult to properly power a study for the identification of *trans*-eQTLs. We show that co-accessibility data can be practically used to reduce the multiple testing burden faced by genetic association studies due to the large search space for *trans* effects. By leveraging co-accessible information, we were able to test single variants against single sites, enabling the identification of tens of thousands of *trans*-caQTLs from only 134 individuals. Further, these analyses translated to gene expression, with 9 *trans*-caQTLs also being *trans*-eQTLs for iPSC relevant genes. These 9 eGenes were identified only examining intra-chromosomal *trans* effects, and only in one tissue, compared to GTEx which had 94 *trans* eGenes across ~50 tissues. Future studies could perform ATAC-seq and RNA-seq in the same individuals to define co-accessibility networks, and then use them to direct the identification *trans*-eQTLs to increase statistical power. Despite our high computational power (a 16 node, 512 core compute cluster with 2.25TB of RAM total), we were only able to test for intra-chromosomal co-accessibility due to computational requirements (this process took multiple months of running LMMs) and a relatively small cohort – yet found tens of thousands of *trans*-caQTLs. Future studies with more power and resources could likely utilize inter-chromosomal co-accessibility to define co-accessibility networks across all pairs of chromosomes and enable the identification of even more *trans*-caQTLs and *trans*-eQTLs.

Acknowledgements

This work was supported in part by a California Institute for Regenerative Medicine (CIRM) grant GC1R-06673 and NIH grants HG008118-01, HL107442-05, DK105541-03 and DK112155-01. RNA-seq were performed at the UCSD IGM Genomics Center with support from NIH grant P30CA023100. WWYG was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number F31HL142151. Whole genome sequencing was performed at Human Longevity, Inc.

Author Contributions

Conceptualization, A.C.D, W.W.Y.G., and K.A.F.; Methodology, A.C.D., W.W.Y.G., and E.N.S; Software, W.W.Y.G.; Validation, W.W.Y.G.; Formal Analysis, W.W.Y.G.; Investigation, A.C.D., W.W.Y.G., P.B.; Data Curation, A.C.D., W.W.Y.G., H.M., M.D.; Writing – Original Draft, A.C.D., W.W.Y.G., M.D., and K.A.F.; Visualization, W.W.Y.G.; Supervision, A.C.D, and K.A.F.; Project Administration, A.C.D., and K.A.F.; Funding Acquisition K.A.F.

Conflicts of Interest

The authors have no conflicts of interest.

Methods

Selection of Individuals form iPSCORE

152 iPSC lines from 134 individuals from iPSCORE were selected for ATAC-seq analysis. These 134 individuals are from multiple ethnicities; among them, 82 individuals belong to 26 families and 52 are unrelated. For all 152 lines, ATAC libraries were generated from matched iPSC and iPSC-derived cardiomyocytes (cardiomyocytes not part of this manuscript).

ATAC-seq

We performed ATAC-seq on 152 iPSC samples using the protocol from Buenrostro et al. (Buenrostro et al., 2013) with small modifications. Frozen nuclear pellets of 2.5 x 10^4 PSCs were thawed on ice and tagmented in total volume of 25µl in permeabilization buffer containing digitonin and 2.5µl of Tn5 from Nextera DNA Library Preparation Kit (Illumina) for 45-75min at 37°C in a thermomixer (500 RPM shaking). To eliminate confounding effects due to index hopping, all libraries within a pool were indexed with unique i7 and i5 barcodes. Libraries were amplified for 12 cycles using NEBNext® High-Fidelity 2X PCR Master Mix (NEB) in total volume of 25µl in the presence of 800nM of barcoded primers (400nM each) from custom synthesized by Integrated DNA Technologies (IDT). Each library was independently sequenced twice on Illumina HiSeq 4000 with paired-end 150bp reads.

ATAC Peak Calling

Peaks were called using MACS2 v2.1.1.20160309[72] with the settings: --nomodel --nolambda --keep-dup all -f BAMPE -g hs. Peaks were called either individually, or

simultaneously on all samples by providing each input sample to MACS2 at the same time with the -t option.

Identification and QC of ATAC-seq peaks

To assess the quality of each sample, we identified and characterized peaks per sample. We first aligned the two sequencing runs for each of the 152 samples individually (304 BAM files), removed duplicates, and to ensure that identified peaks represented TF binding sites rather than Tn5 insert sites flanking nucleosomes, filtered to read inserts ≤140bp in length. Following this processing, we separately called peaks on each of these 304 BAM files using MACS2. To assess the quality of each sample, we examined the fraction of reads in peaks and the percent of peaks falling within active regions of the genome as defined by ROADMAP chromatin states 1,2,3,5,6,7, and 11 for iPSC (E020). We found the mean FRiP of the 304 samples to be 13%, and the mean percent of peaks in active regions to be 50%. We merged the two fastq files for each sample, re-removed duplicates, and re-filtered to read inserts ≤140bp in length, producing the final set of 152 BAM files for downstream analyses.

Determining sample coverages

Coverages were obtained using the featureCount package from subread v1.5.0[78] for each sample individually on a set of peaks or summits. Next, counts were TMM-

normalized using edgeR v3.12.1[52] for each peak call or summit call set across all individuals.

Creating a reference set of ATAC-seq summits across the 152 samples

ATAC-seq identifies punctate regions of accessible chromatin which demarcate transcription factor (**TF**) binding sites. However, as peaks must be called from the data *de novo*, rather than identified from a set of known sites as per gene expression analyses, it is important to identify a consistent "reference" set of ATAC-peaks that can be compared consistently across samples. We thus called peaks and summits using MACS2 on all samples simultaneously, identifying a total of 859,563 peaks with 1,839,425 summits. We sought to filter out broad regions with low coverage while maintaining peaks that potentially contained multiple real TF binding sites. We started by finding the normalized coverage in TMMs for each sample across all ~1.8M summits. We found the vast majority of summits to have a median of ≤ 2.0 TMMs across the 152 samples. As the summits represent the high points in the peaks, we filtered the peaks based on the median coverage of the contained summits (medCov). Specifically, we tested three filters based on the maximum medCov of all summits within each peak: maximum median coverage of any of their contained summits 0x, 1x, or 2x the medianmax(medCov) across all peaks. The 0-median filtered set of peaks contained 859,563 peaks with a similar size distribution (notably, this step filtered some peaks as they had a median of 0 TMM across individuals at all of their summits, indicating most individuals did not have the peak); the 1-median filtered set of peaks contained 546,476 peaks with approximately two-fold enrichment in smaller peaks; and the 2-median filtered set contained 187,046 peaks with

approximately 5-fold enrichment at smaller peaks. To determine if this filtering process removed expected true ATAC peaks, we examined the chromatin states each summit lied in, and performed a Fisher's exact test on the remaining summits vs the filtered summits. We found that, across all filtering, the remaining peaks were enriched for promoters (up to 8-fold), enhancers (up to 3-fold), and bivalent chromatin (up to 22-fold), with larger enhancer enrichment in the 0- and 1- median filtered sets compared to the 2-median filtered set, and the inverse for promoters. Due to the large number of peaks removed by the 2-median filter, the enrichment of small peaks in the 1-median filter, and similar chromatin enrichment profiles (with 1-median filtering leaning toward enhancer enrichment), we chose to filter the peaks with the 1-median threshold, producing a final set of 546,476 peaks with a median size of 221bp, and 1,215,376 summits for analyses.

H3K27AC ChIP-seq experiments and peak calling

We performed H3K27AC ChIP-seq in 52 iPSC samples from 46 lines from 36 individuals from iPSCORE. Pellets of formaldehyde -crosslinked iPSCs were lysed and sonicated in 110 µl of SDS Lysis Buffer (0.5% SDS, 50mM Tris-HCl pH 8.0, 20mM EDTA, 1x cOmplete™ Protease Inhibitor Cocktail (Sigma)) using Diagenode Bioruptor UCD-200 (Diagenode) or Covaris E220 Focused-ultrasonicators (Covaris). For each sample, 1 µg of H3K27ac antibody (Abcam ab4729) was coupled for 2-4h to 11µl of Protein G Dynabeads (Thermo Scientific) and used for overnight chromatin immunoprecipitation in IP buffer (1% Triton X-100, 0.1% DOC, 1x TE buffer, 1x cOmplete™ Protease Inhibitor Cocktail). 40-45µg of chromatin (66ug – 1 sample) was used for immunoprecipitation. Beads with immunoprecipitated chromatin were washed

five times with 150 µl of RIPA buffer (50mM HEPES pH 8.0, 1% NP-40, 0.7% DOC, 500mM LiCl, 1mM EDTA, 1x cOmplete™ Protease Inhibitor Cocktail) and once with 1X TE buffer (10mM Tris-HCl pH 8.0, 1mM EDTA). Next samples were eluted in 150 µl of ChIP Elution Buffer (1% SDS, 10mM Tris-HCl pH 8.0, 1mM EDTA) and reverse crosslinked by incubation for overnight at 65˚C and subsequent incubation with 5 µl RNAse (Sigma) for 1h at 37˚C and Proteinase K Solution (20 mg/mL, Thermo Fisher Scientific) for 1h at 55˚C. After reverse crosslinking, samples were purified with MIniElute PCR purification kit (Qiagen) or with DNA Clean & Concentrator kit (Zymo), eluted in 25µl of EB buffer (Qiagen)and Qubit (Thermo Scientific) quantified. Libraries were generated using KAPA Hyper Prep Kit (KAPA Biosystems) and KAPA Real Time Library Amplification Kit (KAPA Biosystems) at Institute for Genomic Medicine at University of California, San Diego. Libraries were barcoded using TruSeq RNA Indexes (Illumina). Libraries were sequenced on an Illumina HiSeq 4000 100bp Paired-End reads. Peaks were called using MACS2 --broadpeak on all samples simultaneously.

ATAC-seq tag distribution at H3K27ac peaks

MakeTagDirectory.pl from HOMER v4.7[115] was used on each sample individually at the set of H3K72ac peaks called on all samples simultaneously. After creating tag directories, annotatePeaks.pl from HOMER was used with -size 1000 -hist 50 -d to find mean coverages per sample at H3K27ac peaks. The average of this coverage was then calculated for plotting.

Chromatin state enrichment at ATAC-seq peaks

To measure the enrichment for particular chromatin states at peaks, we used bedtools[5] to identify the number of base pairs present in each chromatin state within ATAC peaks, and compared this proportion to that of the coverage of each state in the entire genome via a Fisher's Exact test.

Enriched transcription factors at accessible sites

To calculate the enrichment for transcription factors for sites, findMotifsGenome.pl from HOMER was used on hg19 with -size 200.

Transcription factor motif prediction

To identify transcription factor (TF) binding sites at sites, FIMO from MEME v4.12.0[116] was used on the 200bp flanking each sites with transcription factors from HOCOMOCO individually. Results were filtered to $q<0.05$ for each TF.

Identifying co-accessible sites

To identify co-accessible sites, we utilized a Linear Mixed-effects Model (LMM) to control for fixed covariate effects, and random effects from kinship, as iPSCORE contains related individuals. First, we quantile normalized TMMs for each accessible site across individuals to remove outlier effects. Next, we utilized Limix v1.0.17 (github.com/limix/limix) and included age, iPSC passage number, sex, and the top 20 PCs from ancestry (previously calculated in Panopolous et. al[38]) as fixed effect covariates, and kinship as a random effect. Kinship values were obtained from DeBoever et. al[42]. We then ran Limix with these covariates on all pairs of sites for each

chromosome, and Benjamini-Hochberg FDR corrected the regression p-values within each chromosome using statsmodels in python, using a significance threshold of $q < 0.05$. To examine within-peak correlation, we used the p-values from these analyses.

Co-accessibility enrichment at co-binding TFs and chromatin looping

To measure co-accessible enrichments for TFs and chromatin looping, pgl files were created by pairing together EZH2 and Suz12 peaks within 500bp of one another, and pgltools[75] was used to convert calls from CTCF ChIA-PET[16] in GM12878 and pHiC in iPSCs[76] to the pgl format. Next, pgltools intersect1D was used to find accessible sites at opposing anchors of loops, or at Ezh2 Suz12 pairs, and p-values were obtained from the co-accessibility analysis for enrichments.

Annotating accessible sites with TF ChIP-seq, gene promoter, and chromatin states

To label TFs, gene promoters, and chromatin states at accessible sites, we utilized all public ChIP-seq from UCSC genome browser in ESC, ROADMAP chromatin state E020 15 state model, and GENCODE promoters. Bedtools was used to identify annotations which overlapped sites, and each site was labelled with all ChIP types, a chromatin state, and a gene (if it overlapped one).

Annotating accessible sites with TF Motifs

Sites were annotated with all motifs they overlapped from the above FIMO analysis using bedtools. Following, sites were filtered for the clustering analysis by: 1) finding the mean TPM of all genes in the 134 individuals; 2) identifying enriched motifs

with HOMER; 3) mapping HOCOMOCCO motif names to GENCODE genes using HOCOMOCCO's metadata information (note: many genes were lost in this process); filtering to TFs whose genes were expressed at $\log_2(TPM) \geq 1$.

Creation of co-accessibility networks

To create co-accessibility networks, we utilized the networkX package for python (v2.1). Edges were added to the network between each FDR $q < 0.05$ co-accessible sites with weights equal to their regression coefficient. All sites within the network were then annotated with chromatin states, gene promoters, TF ChIP-seq, and TF motifs, and edges were annotated with the genomic distance between sites. Following, all networks were combined into a single network for ease of use. Chromosome networks are induced by taking the subset of nodes within a given chromosome, and networks centered on a node are induced by subsetting to the node and all its neighbors.

Clustering of annotations in co-accessibility networks using permutation tests

To create null networks, node labels (ie site names) were shuffled 25k times, and all annotations were shuffled with them; edges remained constant. To calculate empirical p-values, for each of the 25k null permutations, the mean edge weight between any two annotations was calculated and compared to the true mean edge weight. Directional empirical p-values were calculated by counting the number of times a stronger mean weight occurred in the null network compared to the original network, using the sign of the true mean edge weight (ie, as 1_Tss and 10_TssBiv had a positive mean weight, we

counted the number of times a larger positive number occurred; as 3_TxFlnk and 15_Quies had a negative mean weight, we counted the number of times a larger negative number occurred). We then calculated an empirical p-value, and signed the p-value by the original mean edge weight sign so that anti-correlations would repel positive correlations for the same TF or chromatin state during clustering (ie TF_A and TF_B positive enrichment should cluster far away from an antagonistic association with TF_C).

Identifying caQTLs and caSites

caQTLs were identified using the qtl_test_lmm function from Limix v1.0.17. TMM data was quantile normalized within each site across individuals. A kinship matrix was included as a random effect to account for relatedness between individuals, and the following variables were included as fixed effect covariates: age, sex, the top 20 principle components for ancestry, and the top 30 PEER factors from the TMM normalized count data (calculated with PEER v1.0[117]). While the ATAC samples were each sequenced twice, as each individual included data from both runs, batch was not included as a covariate. All sites were tested as we previously filtered our site set based on coverage (see above in methods). All SNVs within 100kb upstream or downstream of each site were utilized for testing. As the data space was large (~1 million sites) we chose a conservative correction approach that was computationally fast: from the p-values for each SNV calculated from Limix, the minimum p-value was chosen from each site and Bonferroni corrected for the within-site number of tested SNVs. These Bonferroni adjusted p-values were then FDR-corrected as a whole across all sites, and sites with an FDR q-value < 0.05 were identified as significant caSites.

Identifying allele specific effects at caSites

To identify allele specific effects (ASE), BAMS were remapped with WASP (v0.2.1), following which all heterozygous variants within 200bp upstream or downstream of a site were utilized. At each site, samtools[118] mpileup was utilized to obtain allele counts. To identify ASE, all variants with 10 or more reads were tested for imbalance via a normal approximation to a binomial so that Z scores could subsequently be combined across individuals in a signed manner via Stouffer's method. P values were calculated for each Stouffer Z score, and ASE sites were identified as those with one variant with an FDR $q < 0.05$.

Concordance in ASE across co-accessible sites

To determine if ASE was similar across co-accessible sites, each node in the co-accessibility network was labelled with its ASE Z score. Next, we identified all pairs of Z scores connected by an edge. Finally, we utilized statsmodels.OLS.fromformula to regress the weaker Z scores against the stronger Z scores with a forced intercept of 0 (as no ASE would correspond to no ASE).

Trans caQTL

To identify putative *trans*-caQTLs, we performed targeted association tests by leveraging the co-accessibility networks. For each *cis*-caSite, we tested the lead variant of the site against its co-accessible partners, and the FDR q-corrected all trans tests

simultaneously with the Benjamini-Hochberg method. For these associations, we included sex, age, passage, the top 5 PCs from ancestry, and the top 20 PEER factors from expression as fixed effects, and kinship as a random effect.

Mediator analysis for trans caQTL

To identify *trans* ca-QTLs whose effects were mediated through *cis*-caSites, we calculated the Sobbel p-value for each variant-cis-trans combination that was FDR q< 0.05 from the putative trans analysis. First, the trans height (Y_{trans}) was regressed against the cis-site height (Y_{cis}), using the SNP genotypes (X_{cis}) as a covariate to obtain the mediator effect ($\beta_{mediated}$) and its standard error ($\sigma_{\beta_{mediated}}$):

$$(1)\quad Y_{trans} = \beta_0 + \beta_{snp}X_{cis} + \beta_{mediated}Y_{cis} + covariates + kinship$$

Next, the *cis* association (β_{cis}) and its standard error ($\sigma_{\beta_{cis}}$) was obtained from prior analysis.

The Sobbel p-value was found using the following Z score:

$$Z = \frac{\beta_{mediated} * \beta_{cis}}{\sqrt{\beta^2_{mediated}\sigma^2_{\beta_{cis}} + \beta^2_{cis}\sigma^2_{\beta_{mediated}}}}$$

We utilized Limix to perform the analysis, and included the *cis* genetic effect as a fixed effect covariate in order to obtain the necessary values as output from Limix.

Validation of GTEx trans-QTLs for co-accessibility

To identify if GTEx trans-QTLs had co-accessibility, we identified all sites at the 32 gene promoters, and all sites within 100kb of the reported eQTL variant (the converse of how we define what variants to test as caQTLs). We then tested all promoter sites against all variant sites using Limix and the methods described in the co-accessibility section. P-values were Bonferroni corrected within each eQTL-eGene pair, and then FDR corrected across all eGenes. A threshold of $q < 0.05$ was used for significance.

Trans eQTL

To identify *trans* eQTLs, we tested *trans*-caQTL variants against the gene whose promoter overlapped the *trans*-caSite, following which we FDR corrected all eQTL tests. For these associations, we included sex, age, iPSC passage number, the top 5 PCs from ancestry, and the top 20 PEER factors from expression as fixed effects, and kinship as a random effect.

Chapter 3, in full, has been submitted for publication of the material as it may appear in Nature Genetics 2019, William W. Young Greenwald, Agnieszka D'Antonio-Chronowska, Paola Benaglio, Hiroko Matsui, Erin N. Smith, Matteo D'Antonio, Kelly A. Frazer. The dissertation author was one of the primary investigators and authors of this paper.

REFERENCES

1 Li, G., Fullwood, M. J., Xu, H., Mulawadi, F. H., Velkov, S., Vega, V., Ariyaratne, P. N., Mohamed, Y. B., Ooi, H. S., Tennakoon, C., Wei, C. L., Ruan, Y. & Sung, W. K. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol* **11**, R22, doi:10.1186/gb-2010-11-2-r22 (2010).

2 Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. & Dekker, J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

3 Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

4 Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. A. & Ren, B. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336, doi:10.1038/nature14222 (2015).

5 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

6 Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., Heard, E., Dekker, J. & Barillot, E. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**, 259, doi:10.1186/s13059-015-0831-x (2015).

7 Sauria, M. E., Phillips-Cremins, J. E., Corces, V. G. & Taylor, J. HiFive: a tool suite for easy and efficient HiC and 5C data analysis. *Genome Biol* **16**, 237, doi:10.1186/s13059-015-0806-y (2015).

8 Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S. & Aiden, E. L. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98, doi:10.1016/j.cels.2016.07.002 (2016).

9 Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res* **24**, 999-1011, doi:10.1101/gr.160374.113 (2014).

10 Harmston, N., Ing-Simmons, E., Perry, M., Baresic, A. & Lenhard, B. GenomicInteractions: An R/Bioconductor package for manipulating and investigating chromatin interaction data. *BMC Genomics* **16**, 963, doi:10.1186/s12864-015-2140-x (2015).

11 Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. The human genome browser at UCSC. *Genome Res* **12**, 996-1006, doi:10.1101/gr.229102 (2002).

12 Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S. & Aiden, E. L. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99-101, doi:10.1016/j.cels.2015.07.012 (2016).

13 Duggal, G., Wang, H. & Kingsford, C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic Acids Res* **42**, 87-96, doi:10.1093/nar/gkt857 (2014).

14 Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., Greenside, P., Srivas, R., Phanstiel, D. H., Pekowska, A., Heidari, N., Euskirchen, G., Huber, W., Pritchard, J. K., Bustamante, C. D., Steinmetz, L. M., Kundaje, A. & Snyder, M. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051-1065, doi:10.1016/j.cell.2015.07.048 (2015).

15 Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C. A., Schmitt, A. D., Espinoza, C. A. & Ren, B. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294, doi:10.1038/nature12644 (2013).

16 Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S. Z., Penrad-Mobayed, M., Sachs, L. M., Ruan, X., Wei, C. L., Liu, E. T., Wilczynski, G. M., Plewczynski, D., Li, G. & Ruan, Y. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**, 1611-1627, doi:10.1016/j.cell.2015.11.024 (2015).

17 Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B. M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S. W., Dimitrova, E., Dimond, A., Edelman, L. B., Elderkin, S., Tabbada, K., Darbo, E., Andrews, S., Herman, B., Higgs, A., LeProust, E., Osborne, C. S., Mitchell, J. A., Luscombe, N. M. & Fraser, P. The pluripotent regulatory circuitry connecting

promoters to their long-range interacting elements. *Genome Res* **25**, 582-597, doi:10.1101/gr.185272.114 (2015).

18 Bickmore, W. A. & van Steensel, B. Genome architecture: domain organization of interphase chromosomes. *Cell* **152**, 1270-1284, doi:10.1016/j.cell.2013.02.001 (2013).

19 Sexton, T. & Cavalli, G. The role of chromosome domains in shaping the functional genome. *Cell* **160**, 1049-1059, doi:10.1016/j.cell.2015.02.040 (2015).

20 Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

21 Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Mol Cell* **62**, 668-680, doi:10.1016/j.molcel.2016.05.018 (2016).

22 Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat Rev Genet* **17**, 772, doi:10.1038/nrg.2016.147 (2016).

23 Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164**, 1110-1121, doi:10.1016/j.cell.2016.02.007 (2016).

24 Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292-301, doi:10.1038/35066075 (2001).

25 Sexton, T., Schober, H., Fraser, P. & Gasser, S. M. Gene regulation through nuclear organization. *Nat Struct Mol Biol* **14**, 1049-1055, doi:10.1038/nsmb1324 (2007).

26 Kadauke, S. & Blobel, G. A. Chromatin loops in gene regulation. *Biochim Biophys Acta* **1789**, 17-25, doi:10.1016/j.bbagrm.2008.07.002 (2009).

27 Won, H., de la Torre-Ubieta, L., Stein, J. L., Parikshak, N. N., Huang, J., Opland, C. K., Gandal, M. J., Sutton, G. J., Hormozdiari, F., Lu, D., Lee, C., Eskin, E., Voineagu, I., Ernst, J. & Geschwind, D. H. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523-527, doi:10.1038/nature19847 (2016).

28 Krijger, P. H. & de Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat Rev Mol Cell Biol* **17**, 771-782, doi:10.1038/nrm.2016.138 (2016).

29 Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K.,

Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R. & Stamatoyannopoulos, J. A. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195, doi:10.1126/science.1222794 (2012).

30 Lupianez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A. & Mundlos, S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012-1025, doi:10.1016/j.cell.2015.04.004 (2015).

31 Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., Suva, M. L. & Bernstein, B. E. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110-114, doi:10.1038/nature16490 (2016).

32 Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A. L., Bak, R. O., Li, C. H., Goldmann, J., Lajoie, B. R., Fan, Z. P., Sigova, A. A., Reddy, J., Borges-Rivera, D., Lee, T. I., Jaenisch, R., Porteus, M. H., Dekker, J. & Young, R. A. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454-1458, doi:10.1126/science.aad9024 (2016).

33 Maass, P. G., Barutcu, A. R., Weiner, C. L. & Rinn, J. L. Inter-chromosomal Contact Properties in Live-Cell Imaging and in Hi-C. *Mol Cell* **70**, 188-189, doi:10.1016/j.molcel.2018.03.021 (2018).

34 Fudenberg, G. & Imakaev, M. FISH-ing for captured contacts: towards reconciling FISH and 3C. *Nat Methods* **14**, 673-678, doi:10.1038/nmeth.4329 (2017).

35 Wang, S., Su, J. H., Beliveau, B. J., Bintu, B., Moffitt, J. R., Wu, C. T. & Zhuang, X. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598-602, doi:10.1126/science.aaf8084 (2016).

36 Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., Leeb, M., Wohlfahrt, K. J., Boucher, W., O'Shaughnessy-Kirwan, A., Cramard, J., Faure, A. J., Ralser, M., Blanco, E., Morey, L., Sanso, M., Palayret, M. G. S., Lehner, B., Di Croce, L., Wutz, A., Hendrich, B., Klenerman, D. & Laue, E. D. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59-64, doi:10.1038/nature21429 (2017).

37 Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., Nguyen, M. L., Rubin, A. J., Granja, J. M., Kazane, K. R., Wei, Y., Nguyen, T., Greenside, P. G., Corces, M. R., Tycko, J., Simeonov, D. R., Suliman, N., Li, R., Xu, J., Flynn, R. A., Kundaje, A., Khavari, P. A., Marson, A., Corn, J. E., Quertermous, T., Greenleaf, W. J. & Chang, H. Y. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet* **49**, 1602-1612, doi:10.1038/ng.3963 (2017).

38 Panopoulos, A. D., D'Antonio, M., Benaglio, P., Williams, R., Hashem, S. I., Schuldt, B. M., DeBoever, C., Arias, A. D., Garcia, M., Nelson, B. C., Harismendy, O., Jakubosky, D. A., Donovan, M. K. R., Greenwald, W. W., Farnam, K., Cook, M., Borja, V., Miller, C. A., Grinstein, J. D., Drees, F., Okubo, J., Diffenderfer, K. E., Hishida, Y., Modesto, V., Dargitz, C. T., Feiring, R., Zhao, C., Aguirre, A., McGarry, T. J., Matsui, H., Li, H., Reyna, J., Rao, F., O'Connor, D. T., Yeo, G. W., Evans, S. M., Chi, N. C., Jepsen, K., Nariai, N., Muller, F. J., Goldstein, L. S. B., Izpisua Belmonte, J. C., Adler, E., Loring, J. F., Berggren, W. T., D'Antonio-Chronowska, A., Smith, E. N. & Frazer, K. A. iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. *Stem Cell Reports* **8**, 1086-1100, doi:10.1016/j.stemcr.2017.03.012 (2017).

39 Ban, H., Nishishita, N., Fusaki, N., Tabata, T., Saeki, K., Shikamura, M., Takada, N., Inoue, M., Hasegawa, M., Kawamata, S. & Nishikawa, S. Efficient generation of transgene-free human induced pluripotent stem cells (iPSCs) by temperature-sensitive Sendai virus vectors. *Proc Natl Acad Sci U S A* **108**, 14234-14239, doi:10.1073/pnas.1103509108 (2011).

40 Lian, X., Zhang, J., Azarin, S. M., Zhu, K., Hazeltine, L. B., Bao, X., Hsiao, C., Kamp, T. J. & Palecek, S. P. Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/beta-catenin signaling under fully defined conditions. *Nat Protoc* **8**, 162-175, doi:10.1038/nprot.2012.150 (2013).

41 Benaglio, P., D'Antonio-Chronowska, A., Greenwald, W. W., DeBoever, C., Li, H., Drees, F., Singhal, S., Matsui, H., D'Antonio, M., Smith, E. N. & Frazer, K. A. Allele-specific NKX2-5 binding underlies multiple genetic associations with human EKG traits. *bioRxiv*, doi:10.1101/351411 (2018).

42 DeBoever, C., Li, H., Jakubosky, D., Benaglio, P., Reyna, J., Olson, K. M., Huang, H., Biggs, W., Sandoval, E., D'Antonio, M., Jepsen, K., Matsui, H., Arias, A., Ren, B., Nariai, N., Smith, E. N., D'Antonio-Chronowska, A., Farley, E. K. & Frazer, K. A. Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. *Cell Stem Cell* **20**, 533-546 e537, doi:10.1016/j.stem.2017.03.009 (2017).

43 Yang, T., Zhang, F., Yardimci, G. G., Song, F., Hardison, R. C., Noble, W. S., Yue, F. & Li, Q. HiCRep: assessing the reproducibility of Hi-C data using a

stratum-adjusted correlation coefficient. *Genome Res* **27**, 1939-1949, doi:10.1101/gr.220640.117 (2017).

44 Forcato, M., Nicoletti, C., Pal, K., Livi, C. M., Ferrari, F. & Bicciato, S. Comparison of computational methods for Hi-C data analysis. *Nat Methods* **14**, 679-685, doi:10.1038/nmeth.4325 (2017).

45 Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J. P., Tanay, A. & Cavalli, G. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572 e524, doi:10.1016/j.cell.2017.09.043 (2017).

46 Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jorgensen, H. F., John, R. M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M. & Fisher, A. G. Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* **8**, 532-538, doi:10.1038/ncb1403 (2006).

47 Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L. & Lander, E. S. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326, doi:10.1016/j.cell.2006.02.041 (2006).

48 Mas, G., Blanco, E., Ballare, C., Sanso, M., Spill, Y. G., Hu, D., Aoi, Y., Le Dily, F., Shilatifard, A., Marti-Renom, M. A. & Di Croce, L. Promoter bivalency favors an open chromatin architecture in embryonic stem cells. *Nat Genet* **50**, 1452-1462, doi:10.1038/s41588-018-0218-5 (2018).

49 Freire-Pritchett, P., Schoenfelder, S., Varnai, C., Wingett, S. W., Cairns, J., Collier, A. J., Garcia-Vilchez, R., Furlan-Magaril, M., Osborne, C. S., Fraser, P., Rugg-Gunn, P. J. & Spivakov, M. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *Elife* **6**, doi:10.7554/eLife.21926 (2017).

50 Lun, A. T. & Smyth, G. K. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16**, 258, doi:10.1186/s12859-015-0683-0 (2015).

51 Lareau, C. A. & Aryee, M. J. diffloop: a computational framework for identifying and analyzing differential DNA loops from sequencing data. *Bioinformatics* **34**, 672-674, doi:10.1093/bioinformatics/btx623 (2018).

52 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

53 Siersbaek, R., Madsen, J. G. S., Javierre, B. M., Nielsen, R., Bagge, E. K., Cairns, J., Wingett, S. W., Traynor, S., Spivakov, M., Fraser, P. & Mandrup, S. Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation. *Mol Cell* **66**, 420-435 e425, doi:10.1016/j.molcel.2017.04.010 (2017).

54 Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**, 1059-1065, doi:10.1038/ng.947 (2011).

55 Mayba, O., Gilbert, H. N., Liu, J., Haverty, P. M., Jhunjhunwala, S., Jiang, Z., Watanabe, C. & Zhang, Z. MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol* **15**, 405, doi:10.1186/s13059-014-0405-3 (2014).

56 Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schopflin, R., Kraft, K., Kempfer, R., Jerkovic, I., Chan, W. L., Spielmann, M., Timmermann, B., Wittler, L., Kurth, I., Cambiaso, P., Zuffardi, O., Houge, G., Lambie, L., Brancati, F., Pombo, A., Vingron, M., Spitz, F. & Mundlos, S. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265-269, doi:10.1038/nature19800 (2016).

57 Paulsen, J., Sekelja, M., Oldenburg, A. R., Barateau, A., Briand, N., Delbarre, E., Shah, A., Sorensen, A. L., Vigouroux, C., Buendia, B. & Collas, P. Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome Biol* **18**, 21, doi:10.1186/s13059-016-1146-2 (2017).

58 Rieber, L. & Mahony, S. miniMDS: 3D structural inference from high-resolution Hi-C data. *Bioinformatics* **33**, i261-i266, doi:10.1093/bioinformatics/btx271 (2017).

59 Lesne, A., Riposo, J., Roger, P., Cournac, A. & Mozziconacci, J. 3D genome reconstruction from chromosomal contacts. *Nat Methods* **11**, 1141-1143, doi:10.1038/nmeth.3104 (2014).

60 Zhu, G., Deng, W., Hu, H., Ma, R., Zhang, S., Yang, J., Peng, J., Kaplan, T. & Zeng, J. Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic Acids Res* **46**, e50, doi:10.1093/nar/gky065 (2018).

61 Hug, C. B., Grimaldi, A. G., Kruse, K. & Vaquerizas, J. M. Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell* **169**, 216-228 e219, doi:10.1016/j.cell.2017.03.024 (2017).

62 Ghavi-Helm, Y., Klein, F. A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W. & Furlong, E. E. Enhancer loops appear stable during development and are

associated with paused polymerase. *Nature* **512**, 96-100, doi:10.1038/nature13417 (2014).

63 Rubin, A. J., Barajas, B. C., Furlan-Magaril, M., Lopez-Pajares, V., Mumbach, M. R., Howard, I., Kim, D. S., Boxer, L. D., Cairns, J., Spivakov, M., Wingett, S. W., Shi, M., Zhao, Z., Greenleaf, W. J., Kundaje, A., Snyder, M., Chang, H. Y., Fraser, P. & Khavari, P. A. Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. *Nat Genet* **49**, 1522-1528, doi:10.1038/ng.3935 (2017).

64 Chambers, E. V., Bickmore, W. A. & Semple, C. A. Divergence of mammalian higher order chromatin structure is associated with developmental loci. *PLoS Comput Biol* **9**, e1003017, doi:10.1371/journal.pcbi.1003017 (2013).

65 Morgan, S. L., Mariano, N. C., Bermudez, A., Arruda, N. L., Wu, F., Luo, Y., Shankar, G., Jia, L., Chen, H., Hu, J. F., Hoffman, A. R., Huang, C. C., Pitteri, S. J. & Wang, K. C. Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping. *Nat Commun* **8**, 15993, doi:10.1038/ncomms15993 (2017).

66 Tan-Wong, S. M., Zaugg, J. B., Camblong, J., Xu, Z., Zhang, D. W., Mischo, H. E., Ansari, A. Z., Luscombe, N. M., Steinmetz, L. M. & Proudfoot, N. J. Gene loops enhance transcriptional directionality. *Science* **338**, 671-675, doi:10.1126/science.1224350 (2012).

67 Tohyama, S., Hattori, F., Sano, M., Hishiki, T., Nagahata, Y., Matsuura, T., Hashimoto, H., Suzuki, T., Yamashita, H., Satoh, Y., Egashira, T., Seki, T., Muraoka, N., Yamakawa, H., Ohgino, Y., Tanaka, T., Yoichi, M., Yuasa, S., Murata, M., Suematsu, M. & Fukuda, K. Distinct metabolic flow enables large-scale purification of mouse and human pluripotent stem cell-derived cardiomyocytes. *Cell Stem Cell* **12**, 127-137, doi:10.1016/j.stem.2012.09.013 (2013).

68 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

69 Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigo, R. & Hubbard, T. J. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).

70 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323 (2011).

71 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv e-prints* **1303** (2013). <http://adsabs.harvard.edu/abs/2013arXiv1303.3997L>.

72 Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

73 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).

74 Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *Ima J Numer Anal* **33**, 1029-1047, doi:10.1093/imanum/drs019 (2013).

75 Greenwald, W. W., Li, H., Smith, E. N., Benaglio, P., Nariai, N. & Frazer, K. A. Pgltools: a genomic arithmetic tool suite for manipulation of Hi-C peak and other chromatin interaction data. *BMC Bioinformatics* **18**, 207, doi:10.1186/s12859-017-1621-0 (2017).

76 Montefiori, L. E., Sobreira, D. R., Sakabe, N. J., Aneas, I., Joslin, A. C., Hansen, G. T., Bozek, G., Moskowitz, I. P., McNally, E. M. & Nobrega, M. A. A promoter interaction map for cardiovascular disease genetics. *Elife* **7**, doi:10.7554/eLife.35788 (2018).

77 Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y. C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K. H., Feizi, S., Karlic, R., Kim, A. R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L. H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J.

R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T. & Kellis, M. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

78 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).

79 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

80 Selvaraj, S., J, R. D., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* **31**, 1111-1118, doi:10.1038/nbt.2728 (2013).

81 van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**, 1061-1063, doi:10.1038/nmeth.3582 (2015).

82 Greenwald, W. W., Chiou, J., Nariai, N., Wang, A., Drees, F., Van de Bunt, M., Ren, B., Sander, M., Frazer, K. A. & Gaulton, K. J. Integrative genetic fine-mapping using epigenome and gene expression data resolves islet regulatory variants influencing type 2 diabetes risk. (In Process).

83 Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat Genet* **48**, 206-213, doi:10.1038/ng.3467 (2016).

84 Consortium, G. T., Laboratory, D. A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G. g., Fund, N. I. H. C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, Biospecimen Collection Source Site, N., Biospecimen Collection Source Site, R., Biospecimen Core Resource, V., Brain Bank Repository-University of Miami Brain Endowment, B., Leidos Biomedical-Project, M., Study, E., Genome Browser Data, I., Visualization, E. B. I., Genome Browser Data, I., Visualization-Ucsc Genomics Institute, U. o. C. S. C., Lead, a., Laboratory, D. A., Coordinating, C., management, N. I. H. p., Biospecimen, c., Pathology, e, Q. T. L. m. w. g., Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).

85 Gate, R. E., Cheng, C. S., Aiden, A. P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M. G., Subramaniam, M., Shamim, M., Hougen, K. L., Wortman, I., Huang, S. C., Durand, N. C., Feng, T., De Jager, P. L., Chang, H. Y., Aiden, E. L., Benoist, C., Beer, M. A., Ye, C. J. & Regev, A. Genetic

determinants of co-accessible chromatin regions in activated T cells across humans. *Nat Genet* **50**, 1140-1150, doi:10.1038/s41588-018-0156-2 (2018).

86 Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J. & Trapnell, C. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell* **71**, 858-871 e858, doi:10.1016/j.molcel.2018.06.044 (2018).

87 Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57**, 159-197, doi:10.1146/annurev.bi.57.070188.001111 (1988).

88 Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B. K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E. & Stamatoyannopoulos, J. A. The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232 (2012).

89 Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**, 21 29 21-29, doi:10.1002/0471142727.mb2129s109 (2015).

90 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).

91 Genomes Project, C., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. & McVean, G. A. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

92 Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., Maurano, M. T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M., Roach, V., Dunn, D., Neri, J., Schafer, A., Hansen, R. S., Kutyavin, T., Giste, E., Weaver, M., Canfield, T., Sabo, P., Zhang, M., Balasundaram, G., Byron, R., MacCoss, M. J., Akey, J. M., Bender, M. A., Groudine, M., Kaul, R. & Stamatoyannopoulos, J. A. An expansive human regulatory lexicon encoded in

transcription factor footprints. *Nature* **489**, 83-90, doi:10.1038/nature11212 (2012).

93 Gusmao, E. G., Allhoff, M., Zenke, M. & Costa, I. G. Analysis of computational footprinting methods for DNase sequencing experiments. *Nat Methods* **13**, 303-309, doi:10.1038/nmeth.3772 (2016).

94 Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y. & Pritchard, J. K. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* **21**, 447-455, doi:10.1101/gr.112623.110 (2011).

95 Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat Genet*, doi:10.1038/s41588-018-0278-6 (2018).

96 Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186, doi:10.1016/j.cell.2017.05.038 (2017).

97 Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *bioRxiv*, 425108, doi:10.1101/425108 (2018).

98 Toenhake, C. G., Fraschka, S. A., Vijayabaskar, M. S., Westhead, D. R., van Heeringen, S. J. & Bartfai, R. Chromatin Accessibility-Based Characterization of the Gene Regulatory Network Underlying Plasmodium falciparum Blood-Stage Development. *Cell Host Microbe* **23**, 557-569 e559, doi:10.1016/j.chom.2018.03.007 (2018).

99 D'Antonio, M., Benaglio, P., Jakubosky, D., Greenwald, W. W., Matsui, H., Donovan, M. K. R., Li, H., Smith, E. N., D'Antonio-Chronowska, A. & Frazer, K. A. Insights into the Mutational Burden of Human Induced Pluripotent Stem Cells from an Integrative Multi-Omics Approach. *Cell Rep* **24**, 883-894, doi:10.1016/j.celrep.2018.06.091 (2018).

100 Panopoulos, A. D., Smith, E. N., Arias, A. D., Shepard, P. J., Hishida, Y., Modesto, V., Diffenderfer, K. E., Conner, C., Biggs, W., Sandoval, E., D'Antonio-Chronowska, A., Berggren, W. T., Izpisua Belmonte, J. C. & Frazer, K. A. Aberrant DNA Methylation in Human iPSCs Associates with MYC-Binding Motifs in a Clone-Specific Manner Independent of Genetics. *Cell Stem Cell* **20**, 505-517 e506, doi:10.1016/j.stem.2017.03.010 (2017).

101 D'Antonio, M., Woodruff, G., Nathanson, J. L., D'Antonio-Chronowska, A., Arias, A., Matsui, H., Williams, R., Herrera, C., Reyna, S. M., Yeo, G. W., Goldstein, L. S. B., Panopoulos, A. D. & Frazer, K. A. High-Throughput and

Cost-Effective Characterization of Induced Pluripotent Stem Cells. *Stem Cell Reports* **8**, 1101-1111, doi:10.1016/j.stemcr.2017.03.011 (2017).

102 Greenwald, W. W., Li, H., Benaglio, P., Jakubosky, D., Matsui, H., Schmitt, A., Selvaraj, S., D'Antonio, M., D'Antonio-Chronowska, A., Smith, E. N. & Frazer, K. A. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nat Commun* **10**, 1054, doi:10.1038/s41467-019-08940-5 (2019).

103 Lukic, S., Nicolas, J. C. & Levine, A. J. The diversity of zinc-finger genes on human chromosome 19 provides an evolutionary mechanism for defense against inherited endogenous retroviruses. *Cell Death Differ* **21**, 381-387, doi:10.1038/cdd.2013.150 (2014).

104 Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Hinrichs, A. S., Gonzalez, J. N., Gibson, D., Diekhans, M., Clawson, H., Casper, J., Barber, G. P., Haussler, D., Kuhn, R. M. & Kent, W. J. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **47**, D853-D858, doi:10.1093/nar/gky1095 (2019).

105 Cao, R. & Zhang, Y. The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. *Curr Opin Genet Dev* **14**, 155-164, doi:10.1016/j.gde.2004.02.001 (2004).

106 Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J. & von Mering, C. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* **45**, D362-D368, doi:10.1093/nar/gkw937 (2017).

107 Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., Garcia Giron, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O. G., Lagarde, J., Martin, F. J., Martinez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M. M., Sycheva, I., Uszczynska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J. S., Gerstein, M., Guigo, R., Hubbard, T. J. P., Kellis, M., Paten, B., Reymond, A., Tress, M. L. & Flicek, P. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773, doi:10.1093/nar/gky955 (2019).

108 Zhao, Y., Xie, X., Liao, W., Zhang, H., Cao, H., Fei, R., Wang, X., Wei, L., Shao, Q. & Chen, H. The transcription factor RFX5 is a transcriptional activator

of the TPP1 gene in hepatocellular carcinoma. *Oncol Rep* **37**, 289-296, doi:10.3892/or.2016.5240 (2017).

109 Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* **48**, 488-496, doi:10.1038/ng.3539 (2016).

110 Weintraub, A. S., Li, C. H., Zamudio, A. V., Sigova, A. A., Hannett, N. M., Day, D. S., Abraham, B. J., Cohen, M. A., Nabet, B., Buckley, D. L., Guo, Y. E., Hnisz, D., Jaenisch, R., Bradner, J. E., Gray, N. S. & Young, R. A. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**, 1573-1588 e1528, doi:10.1016/j.cell.2017.11.008 (2017).

111 Oler, A. J., Alla, R. K., Roberts, D. N., Wong, A., Hollenhorst, P. C., Chandler, K. J., Cassiday, P. A., Nelson, C. A., Hagedorn, C. H., Graves, B. J. & Cairns, B. R. Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat Struct Mol Biol* **17**, 620-628, doi:10.1038/nsmb.1801 (2010).

112 Pierce, B. L., Tong, L., Chen, L. S., Rahaman, R., Argos, M., Jasmine, F., Roy, S., Paul-Brutus, R., Westra, H. J., Franke, L., Esko, T., Zaman, R., Islam, T., Rahman, M., Baron, J. A., Kibriya, M. G. & Ahsan, H. Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. *PLoS Genet* **10**, e1004818, doi:10.1371/journal.pgen.1004818 (2014).

113 Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M. & Lancet, D. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics* **54**, 1 30 31-31 30 33, doi:10.1002/cpbi.5 (2016).

114 Dierickx, P., Vermunt, M. W., Muraro, M. J., Creyghton, M. P., Doevendans, P. A., van Oudenaarden, A., Geijsen, N. & Van Laake, L. W. Circadian networks in human embryonic stem cell-derived cardiomyocytes. *EMBO Rep* **18**, 1199-1212, doi:10.15252/embr.201743897 (2017).

115 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

116 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).

117 Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500-507, doi:10.1038/nprot.2011.457 (2012).

118 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing, S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).