2019

Methods for Joint Normalization and Comparison of Hi-C data

John C. Stansfield
Virginia Commonwealth University

Methods for Joint Normalization and Comparison of Hi-C data

By

John Stansfield

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Biostatistics

Virginia Commonwealth University

Doctoral Committee:

Mikhail G. Dozmorov	Department of Biostatistics (Committee Chair)
Roy T. Sabo	Department of Biostatistics
Nitai Mukhopadhyay	Department of Biostatistics
Vladimir Vladimirov	Virginia Institute for Psychiatric and Behavioral Genetics
Joseph McClay	Department of Pharmacotherapy & Outcomes Science

2019

Acknowledgements

First, I would like to thank my advisor, Dr. Mikhail Dozmorov, for all his mentoring and incredible work ethic. Dr. Dozmorov was always available to answer questions whether by email or in person and his guidance has has helped me to accomplish quite a bit in a short amount of time. I also need to thank my committee members, Dr. Roy Sabo, Dr. Nitai Mukhopadhyay, Dr. Vladimir Vladimirov, and Dr. Joseph McClay, for their input, suggestions, and edits to my dissertation. I'd also like to thank Dr. David Wheeler for doing the work necessary to keep me on the T32 training grant for the past 3 years.

I would like to thank Russell Boyle and Dr. Adam Sima for the Biostatistical Consulting Laboratory (BCL). BCL has provided me with ample opportunities to work on interesting and varied consulting projects. These projects gave me a great experience in working on real problems and I have learned how to better communicate with scientists in other fields.

I'd like to thank my family and especially my parents for their support through graduate school and everything that came before. To Kellen and Sprio, thanks for bouncing ideas around, helping with projects, and all of our extracurricular activities. Thanks to Phil M, Conor K, Jack C, Michael S, Joe P for being good friends and still keeping in touch. To Matt C, thanks for the fun arguments and more esoteric discussions. Finally, I'd like to thank Mike C and all of Richmond Kunst des Fechtens for letting me play with swords in the garage. I had a great time with you all at practices, all the tournaments, and other events we went to.

Table of Contents

Table of Contents

List of Tables

List of Figures

Abstract

Methods for Joint Normalization and Comparison of Hi-C data

By John C. Stansfield

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2019

Major Director: Mikhail G. Dozmorov, Ph.D.,

Assistant Professor, Department of Biostatistics

The development of chromatin conformation capture technology has opened new avenues of study into the 3D structure and function of the genome. Chromatin structure is known to influence gene regulation, and differences in structure are now emerging as a mechanism of regulation between, e.g., cell differentiation and disease vs. normal states. Hi-C sequencing technology now provides a way to study the 3D interactions of the chromatin over the whole genome. However, like all sequencing technologies, Hi-C suffers from several forms of bias stemming from both the technology and the DNA sequence itself. Several normalization methods have been developed for normalizing individual Hi-C datasets, but little work has been done on developing joint normalization methods for comparing two or more Hi-C datasets. To make full use of Hi-C data, joint normalization and statistical comparison techniques are needed to carry out experiments to identify regions where chromatin structure differs between conditions.

We develop methods for the joint normalization and comparison of two Hi-C datasets, which we then extended to more complex experimental designs. Our normalization method is novel in that it makes use of the distance-dependent nature of chromatin interactions. Our modification of the Minus vs. Average (MA) plot to the Minus vs. Distance (MD) plot allows for a nonparametric data-driven normalization technique using loess smoothing. Additionally, we present a simple statistical method using Z-scores for detecting differentially interacting regions between two datasets. Our initial method was published as the Bioconductor R package HiCcompare http://bioconductor.org/packages/HiCcompare/.

We then further extended our normalization and comparison method for use in complex Hi-C experiments with more than two datasets and optional covariates. We extended the normalization method to jointly normalize any number of Hi-C datasets by using a cyclic loess procedure on the MD plot. The cyclic loess normalization technique can remove between dataset biases efficiently and effectively even when several datasets are analyzed at one time. Our comparison method implements a generalized linear model-based approach for comparing complex Hi-C experiments, which may have more than two groups and additional covariates. The extended methods are also available as a Bioconductor R package http://bioconductor.org/packages/multiHiCcompare/. Finally, we demonstrate the use of HiCcompare and multiHiCcompare in several test cases on real data in addition to comparing them to other similar methods (https://doi.org/10.1002/cpbi.76).

Chapter 1: Introduction

1.1 Motivation

Only ~2% of the human genome encodes genetic information used to make proteins, the building blocks of cells. Even more surprising is that the relatively static genomic information gives rise to the observed diversity of tissues and cell types. This diversity is partly explained by the discovery of epigenomic (Greek for *on top of* the genome) modifications such as DNA methylation (Bernstein et al. 2007). In contrast to the static genome sequence, epigenomic modifications are dynamic, with each cell type characterized by a distinct epigenomic signature (Hemberger et al. 2009; Shipony et al. 2014). These epigenomic modifications are associated with gene expression changes, and are currently regarded as a well-established regulatory layer (e.g., an increase of DNA methylation in a gene promoter will typically lead to a decrease in gene expression) (Bird 2002).

The Three-dimensional (3D) chromatin structure of the genome is emerging as a unifying regulatory framework orchestrating gene expression by bringing transcription factors, enhancers and co-activators in spatial proximity to the promoters of genes (Franke et al. 2016; Symmons et al. 2014; Sexton and Cavalli 2015; Li et al. 2012; Papantonis and Cook 2013; Laat and Grosveld 2003; Mora et al. 2016; Mifsud et al. 2015; Shavit and Lio' 2014a; Osborne et al. 2004). Together with epigenomic profiles, changes in chromatin interactions shape cell type-specific gene expression (Fernandez et al. 2012; Liu et al. 2008; Jin et al. 2013a; Lieberman-Aiden et al. 2009a; Sanyal et al. 2012; Schmitt et al. 2016; Nora et al. 2012), as well as misregulation of oncogenes and tumor suppressors in cancer (Taberlay et al. 2016a; Hnisz et al. 2016a; Franke et al. 2016; Lupiáñez et al. 2016a). Identifying changes in chromatin interactions is the next logical step in understanding genomic regulation.

The first sequencing technologies that allowed for the study of the 3D structure of the genome were Chromatin Conformation Capture (3C) methods (Dekker et al. 2002a). 3C methods can only capture the structure of a subset of the genome at a single time. The development of 3C then led to several extensions of the method including 4C and 5C, which allowed for more of the genome to be captured at one time. Finally, the introduction of Hi-C technology by Lieberman-Aiden allowed for the capture of all vs. all long-distance chromatin interactions across the entire genome (Lieberman-Aiden et al. 2009a). Hi-C captures the conformations of the chromosomes by first crosslinking cells with formaldehyde. Then the DNA is digested with a restriction enzyme that leaves a 5'-overhang. The 5'-overhangs are filled with a biotinylated residue, and then the blunt-end fragments are ligated that favor ligation events between cross-linked DNA fragments. The ligated DNA samples produced are the joined fragments of DNA that were in close spatial proximity inside of the nucleus. The junction of the fragments are marked with biotin. A Hi-C library can then be created by shearing the DNA and selecting the fragments containing biotin. The library is then sequenced to produce the raw Hi-C data which can then be further analyzed (Lieberman-Aiden et al. 2009a).

Study of the 3D chromatin structure of the human genome has proven it to be highly organized (Dixon et al. 2012; Rao et al. 2014a) into (Fraser et al. 2015; Sexton et al. 2012) chromosome territories (Cremer and Cremer 2010), topologically associated domains (TADs) (Dixon et al. 2012; Jackson and Pombo 1998; Ma et al. 1998; Nora et al. 2012; Sexton et al. 2012), smaller sub-TADs (Phillips-Cremins and Corces 2013a; Rao et al. 2014a) and, on the most local level, chromatin loops (Rao et al. 2014a; Dowen et al. 2014a; Ji et al. 2016a). These structural units of the chromatin aid in coordinated gene expression (Franke et al. 2016; Symmons et al. 2014; Sexton and Cavalli 2015; Li et al. 2012; Papantonis and Cook 2013; Laat and Grosveld 2003; Mora et al. 2016; Mifsud et al. 2015; Shavit and Lio' 2014a; Osborne et al. 2004; Schoenfelder et al. 2010). The organization of the chromatin plays a role in cell type-specific gene expression (Dowen et al. 2014a; Ji et al. 2016a; Phillips-Cremins and Corces 2013a; Rao et al. 2014a; Vietri Rudan et al. 2015), recombination (Jhunjhunwala et al. 2009) and X chromosome inactivation (Nora et al. 2012; Crane et al. 2015).

Hi-C data allows for the production of chromatin contact maps (Hi-C matrices) in which each cell of the matrix represents a pair of interacting genomic regions. To produce a Hi-C matrix, the genome is divided up into "bins" of a specific size. The number of basepairs in each bin represents the "resolution" of the matrix. Typical resolutions of Hi-C data are 1 megabase (MB), 500 kilobase (KB), 100KB, 50KB, 40KB, 10KB, 1KB, and, most recently achieved, a resolution of 750bp (Bonev et al. 2017). Each bin represents a specific genomic region and the number of times regions are sequenced together is the value in each cell of the matrix. This value is

known as the interaction frequency (IF) for that pair of regions. Higher IFs represent regions that were closer in proximity to each other when sequenced while low or zero value IFs represent regions with low levels of interaction. Figure 1.1 displays a representation of a Hi-C matrix. The cartoon chromosomes on the X and Y-axes indicate the genomic regions in the bins. It is important to note that Hi-C matrices are square and symmetric due to the all vs. all nature of the data.

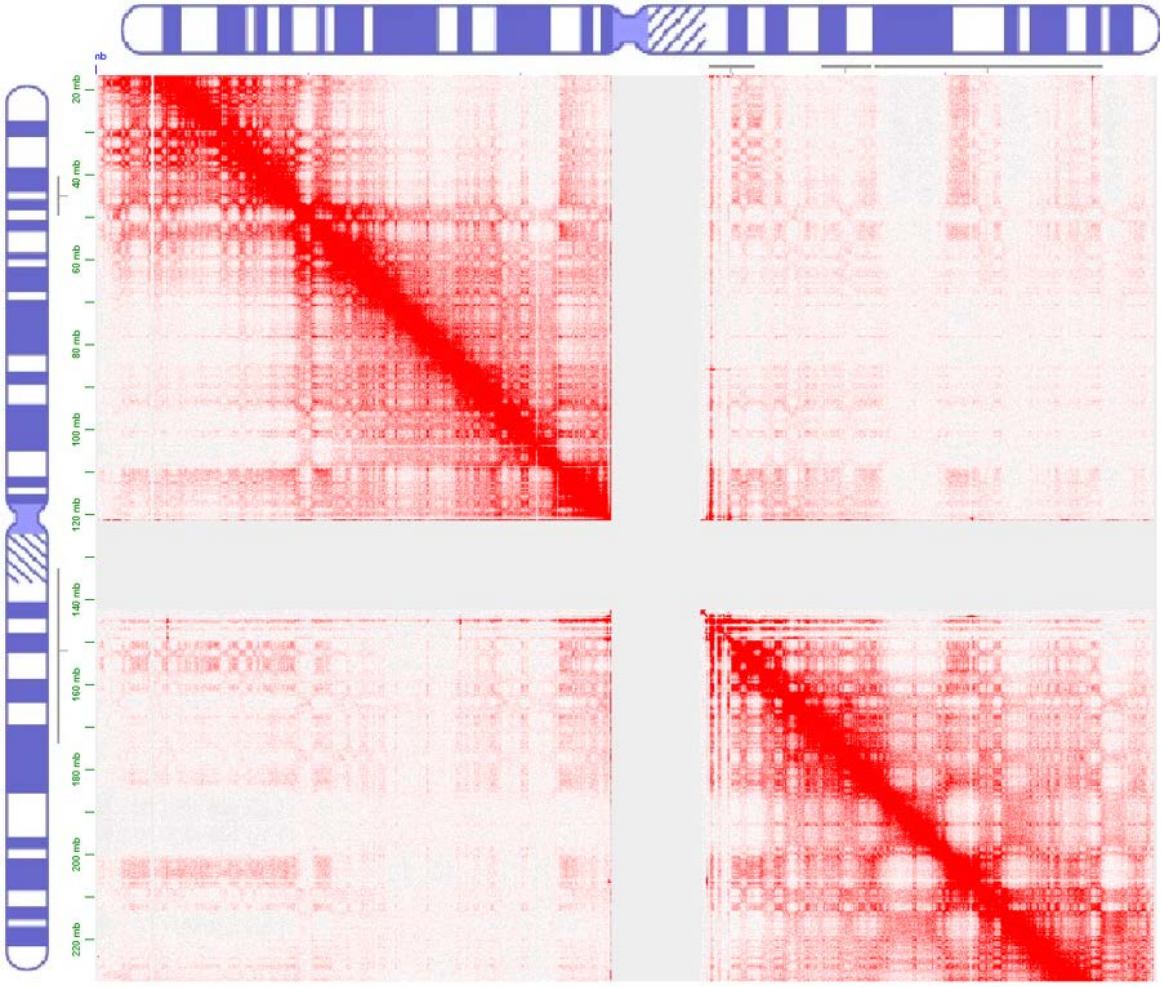

Figure 1.1 Illustration of a chromatin contact map derived from Hi-C data. Each cell of the matrix represents the interaction frequency of a pair of genomic regions.

Hi-C data is strongly distance dependent. Each off-diagonal trace of a Hi-C matrix represents an increase of one unit distance (where unit distance is the resolution of the data) between the pairs of interacting regions. Naturally, interactions occurring at shorter genomic distances are more likely and thus tend to have larger IFs, while interactions occurring at long distance are less likely to occur and correspondingly tend to have smaller IFs. The regions near the diagonal of a Hi-C matrix have the highest intensity of IFs while the corner of the matrix tends to experience much higher sparsity. As distance increases the values of the IFs generally decrease; however, there is a good deal of variation between samples and even chromosomes (Figure 1.2).

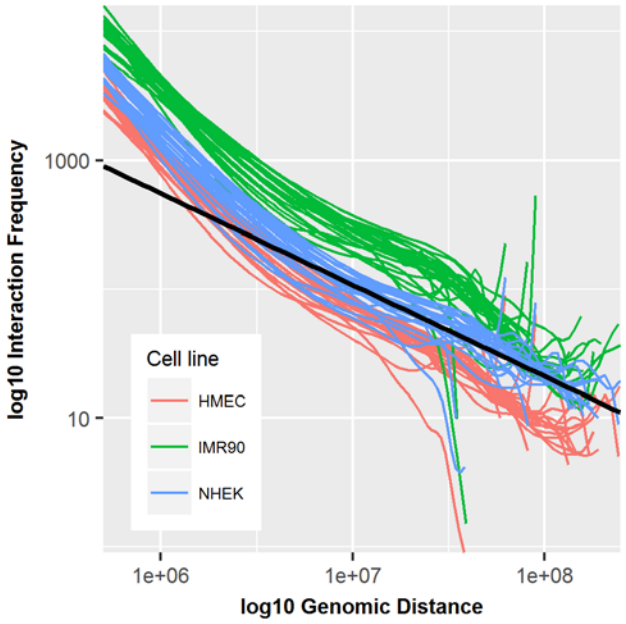

Figure 1.2 Distance-dependent decay of interaction frequencies in Hi-C data. Data from HMEC, IMR90, and NHEK cell lines at 500KB resolution. Black line represents the ideal power-law fit. Each colored line is from a separate chromosome showing that there is variation in the decay between cell lines and chromosomes.

The resolution of Hi-C data also plays a large role in an analysis. Low-resolution data (1MB, 500KB) tend to have much larger IFs due to the sheer size of the genomic bins used while high-resolution data (50KB, 10KB, etc.) exhibit high degrees of sparsity. Because high-resolution data has many more bins, it also means there is a greater range of unit distances, and thus the distance-dependent decay of IFs plays a larger role in high-resolution data. Improvements in sequencing technology have allowed for higher resolution Hi-C data, however much of this data still suffers from sparsity, especially for long-range interactions, which presents many challenges for analysis.

Soon after public Hi-C datasets became available, it was clear that technology- and DNA sequence-driven biases substantially affect chromatin interactions (Yaffe and Tanay 2011a). The technology-specific biases include cutting length of a restriction enzyme (HindIII, MboI, or NcoI), cross-linking conditions, circularization length, etc. (O'Sullivan et al.a; Cournac et al. 2012). The DNA sequence-driven biases include GC content, mappability, nucleotide composition (Yaffe and Tanay 2011a). Discovery of these biases led to the development of methods for normalizing individual datasets (Lieberman-Aiden et al. 2009a; Imakaev et al. 2012a; Yaffe and Tanay 2011a; Knight and Ruiz 2012a). Although normalization of individual datasets improves reproducibility within replicates of Hi-C data (Imakaev et al. 2012a; Yaffe and Tanay 2011a; Hu et al. 2012a), these methods do not consider biases between multiple Hi-C datasets.

Accounting for the between-dataset biases is critical for the correct identification of chromatin interaction changes between, e.g., disease-normal states, or cell types. Left unchecked, biases can be mistaken for biologically relevant differential interactions. While DNA sequence-driven biases affect two datasets similarly (e.g., CG content of genomic regions tested for interaction differences is the same), technology-driven biases are poorly characterized and affect chromatin interactions unpredictably. Importantly, another source of chromatin interaction differences are large-scale genomic rearrangements, such as copy number variations (Harewood et al. 2017; Servant et al. 2015), a frequent event in cancer genomes (Zink et al. 2004; Rickman et al. 2012a). Accounting for such biases is needed for the detection of differential chromatin interactions between Hi-C datasets.

After normalization of biases between datasets, there is a need for methods to perform comparisons. It is of interest to detect differences in the 3D structure of the chromatin between different cell types. Past studies have attempted to compare Hi-C datasets using simple overlap analyses where "significant interactions" within a single Hi-C matrix are overlapped with those from another matrix (Durand et al. 2016a). Other studies have

used correlation between eigen vectors and Euclidean distance of IFs as methods to compare Hi-C matrices (Battulin 2015). To the best of our knowledge, only four methods attempt the comparative analysis of multiple Hi-C datasets. The `diffHic` method is an extension of the general differential expression analysis operating on individual raw sequencing data (Lun and Smyth 2015a). `diffHic` leaves the user with the challenges of sequencing data storage, the computational burden of processing, normalization, summarization, and other bioinformatics heavy lifting. The `HiCCUPS` algorithm (Rao et al. 2014a) detects chromatin interaction "hotspots", chromatin interactions enriched relative to the local background, in individual Hi-C datasets. Hotspots are then compared between datasets by simply overlapping them. This approach does not distinguish "hotspots" detected due to local biases and does not quantify the significance of the differences. One of only two methods to statistically compare processed Hi-C dataset is `ChromoR` (Shavit and Lio' 2014a). However, in our tests, it failed to detect any differential chromatin interactions in real Hi-C data, perhaps due to the use of the parametrically constrained model, an approach that has been criticized (Witten and Noble 2012). The second method for statistical comparison of processed Hi-C data is `FIND`. `FIND` uses a spatial Poisson process to detect differences between two Hi-C experimental conditions (Djekidel et al. 2018a). `FIND` is presented as a tool for high-resolution Hi-C data and treats interactions as spatially dependent on surrounding interactions, but relies on standard individual normalization techniques.

Due to a sparsity of methods, very few differential analyses of Hi-C data have been performed. Of the previously performed differential analyses, the majority are focused on comparing cancer and normal Hi-C datasets, which contain large-scale genomic rearrangements that can easily be detected. Cell- and tissue-specific 3D differences in other cell types and tissues, such as among immune cell types and brain tissues, remain virtually uncharacterized. The brain consists of well-defined anatomical and functional structures each responsible for distinct neurological tasks and represents an interesting topic of study in the field of 3D genomics. However, the study of the 3D chromatin structure of the human brain has been further hampered due to a limited sample availability and ethical considerations related to the collection of brain tissue. Although gene expression and epigenomic programs from different regions of the human brain have been outlined (BrainSpan, CommonMind, NIH Roadmap Epigenomics), the systematic understanding of the regulation mechanism that drives them - the 3D chromatin structure - is lacking. Our proposed research on defining chromatin regions differentially interacting across brain regions will further enrich and complement existing "omics" data on the brain and enable the holistic understanding of the brain's genomics.

1.2 Current methods for Hi-C data processing

1.2.1 Normalization methods

Several normalization methods have been developed for dealing with bias in Hi-C data. Many of these methods are designed to only normalize a single contact map at a time and remove technological or biological dependent biases. These individual normalization methods can help researchers study the interactions within a single cell type or biological condition and have been used in many previous studies. However, these individual methods are not suitable for the comparison of Hi-C datasets, which is necessary to discover differences in the 3D structure of the genome between different cellular conditions.

Common individual normalization methods include the `ChromoR` method (Shavit and Lio' 2014a) which applies the Haar-Fisz Transform (HFT) to decompose a Hi-C contact map. HFT assumes the IFs in the contact map are distributed as a Poisson random variable. After HFT decomposition, wavelet shrinkage methods for Gaussian noise are applied for de-noising. The contact map is then reconstructed with the inverse HFT. The `ChromoR` R package provides the `correctCIM` function to perform this normalization.

`ICE` (iterative correction and eigenvector decomposition) normalization (Imakaev et al. 2012a) functions by modeling the expected IF_{ij} for every pair of regions *(i,j)* as $E_{ij} = B_i B_j T_{ij}$, where B_i and B_j are the biases and T_{ij} is the true matrix of normalized IFs. The maximum likelihood solution for the biases B_i is obtained by iterative correction. It attempts to make all regions equally visible and was shown to perform as well as the explicit bias correction method by Yaffe and Tanay (Belton et al. 2012). `ICE` normalization can be performed using the `HiTC` R package's `normICE` function or as a step of the Hi-C processing pipeline `HiC-Pro`.

KR (Knight-Ruiz) normalization (Knight and Ruiz 2012a) is another "equal visibility" algorithm that balances a square non-negative matrix A by finding a diagonal scaling of A such that $P = D_1 A D_2$ sums to one. The KR algorithm uses an iterative process to find D_1 and D_2 scaling matrices by alternately normalizing columns and rows in a sequence of matrices using an approximation of Newton's method. The KR normalization method was re-implemented in R using the published matlab code (Knight and Ruiz 2012a) and is included in the HiCcompare package as the KRnorm function. KR normalization can also be performed as a function of the Juicer Hi-C processing pipeline.

SCN (Sequential Component Normalization) (Cournac et al. 2012) is a method that is broadly generalizable to many Hi-C experimental protocols. It attempts to smooth out biases due to GC content and circularization. SCN works by first normalizing each column vector of a Hi-C contact matrix to one using the Euclidean norm. Then each row of the resulting matrix is normalized to one using the row Euclidean norm. This process is repeated until convergence (usually 2 to 3 iterations). The SCN method was re-implemented in R and included in the HiCcompare package as the SCN function.

MA (Minus Average normalization) (Lun and Smyth 2015a) is a commonly used joint normalization method for genomic data. It is based on the MA plot (a variant of the Bland-Altman plot) where the data is plotted according to the Average log counts (or counts per million) and the log Minus (difference) between the two data sets. A loess model is then fit to this plot, and the residuals for the fit can be used to smooth the data sets. MA normalization was implemented in R and included in the HiCcompare packages as the MA_norm function. MA normalization is also used within the diffHic R package.

1.2.2 Comparison methods (aka differential analysis)

The diffHic method (Lun and Smyth 2015a) is an extension of edgeR, an R package originally designed for RNA sequencing experiments. diffHic operates on unprocessed Hi-C data in the form of .BAM files. This requires the user to download and convert the raw sequencing output of a Hi-C library in order to make use of the package (a time and storage consuming task) whereas many public Hi-C datasets are available already in a processed contact map format such as the data available from the Aiden lab website or from cooler (http://cooler.readthedocs.org/en/latest/). diffHic uses the generalized linear model framework of edgeR to compare Hi-C data between different cellular conditions.

The HiCCUPS algorithm (Rao et al. 2014a) detects chromatin interaction "hotspots" within a Hi-C contact map. HiCCUPS operates on data in the .hic file format, the final product of the Juicer Hi-C pipeline developed by the Aiden Lab. .hic files are compressed containers for Hi-C contact maps, which can be extracted into a plain text format. HiCCUPS operates on a Hi-C contact map by finding peak "pixels" with higher intensity than the surrounding area. To compare peaks between different datasets, they perform a simple overlap analysis of the significant peaks. Since no joint normalization is performed before HICCUPS, it is possible many of the differences in peaks could be due to biases between datasets. There is also no way to measure the significance of the differences in the peaks between datasets.

chromoR is an R package (Shavit and Lio' 2014a) with functions for normalization and difference detection of processed Hi-C data. Unlike diffHic, chromoR takes its input in the form of processed Hi-C contact maps. This removes the challenges of processing the raw data for the user. chromoR uses a wavelet variance stabilization method for normalizing the data. To detect differences, it uses a wavelet Poisson change point detection algorithm. However, in our tests, it failed to detect differential chromatin interactions in real Hi-C data, perhaps due to the use of the parametrically constrained model, an approach that has been criticized (Witten and Noble 2012).

FIND is an R package (Djekidel et al. 2018a) providing tools for the comparative analysis of Hi-C data. FIND focuses on finding differences between experimental conditions in extremely high-resolution data (1KB - 5KB). FIND treats interactions as spatially dependent on surrounding interactions by using a spatial Poisson process to detect differences. In our testing of FIND, we found that it suffers from very long run times and does not seem to function at all on Hi-C data at resolutions in the range of 100KB - 10KB, perhaps due to increased

sparsity of Hi-C data at these resolutions. Additionally, since most Hi-C data is not sequenced at a deep enough level to support < 5KB resolution contact maps the results of `FIND` are questionable.

1.3 Aims

Our goal is to develop methods for the joint normalization and comparison of Hi-C datasets. First, we will focus on the joint normalization and comparison of two Hi-C datasets. This method will then be applied to a set of Hi-C data for two regions of the human brain. Finally, we will develop methods for the joint normalization and comparison of Hi-C data when there are multiple replicates for each experimental condition. These methods will be developed into R packages that will be freely available for the scientific community to use.

1.3.1 Aim 1: Joint normalization and difference detection for two Hi-C datasets

The production of Hi-C data requires large amounts of money, time, computational power, and storage space. These constraints drastically limit the sample sizes and number of replicates for experiments. Many Hi-C experiments already in the public domain only have a single replicate for each cellular condition. Thus, there is a need for a method to compare Hi-C data between two conditions when only a single replicate is available for each condition. Additionally, when there are replicates available it is common practice to combine (pool) these replicates to produce a single contact map that will have a lower level of sparsity (Won et al. 2016). We jointly normalize Hi-C datasets using loess regression on what we term the MD plot (Minus vs. Distance plot). After normalization, we perform difference detection to identify the pairs of regions that are interacting significantly differently between the experimental conditions. We compare our normalization and difference detection methods to the existing methods. The methods will be compiled into an R package, `HiCcompare` and released on Github and Bioconductor.

1.3.2 Aim 2: Analysis of Brain data with HiCcompare

The human brain is a complex organ composed of distinct anatomical and functional regions and cell types. Being the central organ for human cognition, it is also one of the most difficult organs to study (Birdsill et al. 2011; Popova et al. 2008). While gene expression and epigenomic changes across human brain regions and cell types have been characterized (Kang et al. 2011; Strand et al. 2007; GTEx Consortium 2013), the dynamics of 3D chromatin interactions integrated with “omics” changes remain undefined. We will use `HiCcompare` to detect differences between Hi-C datasets generated from the amygdala and the prefrontal cortex. Brain region-specific gene expression and epigenomic differences (ENCODE, Roadmap, GTeX, etc. data) will be tested for association with the 3D changes using functional enrichment analysis.

1.3.3 Aim 3: Joint normalization and difference detection with replicate Hi-C datasets

As sequencing costs further decrease and Hi-C methods become more refined, it is natural that more and better quality Hi-C data will become available. This will allow for more replicates of experimental conditions to be produced and necessitate a need for methods to make use of them. We will develop a distance-centric cyclic loess normalization method for the joint normalization of multiple Hi-C datasets. Next, we will develop a general linear model (GLM) approach for the differential analysis of replicate Hi-C data. This aim proposes a Bayesian approach for “borrowing” information across Hi-C replicates. These methods will be implemented into an additional R package that will be released on Github and Bioconductor.

Chapter 2: Aim 1 - Joint normalization and difference detection for two Hi-C datasets

2.1 Introduction

The 3D chromatin structure of the genome is emerging as a unifying regulatory framework orchestrating gene expression by bringing transcription factors, enhancers and co-activators in spatial proximity to the promoters of genes (Mifsud et al. 2015; Sexton and Cavalli 2015; Li et al. 2012; Papantonis and Cook 2013). Changes in chromatin interactions shape cell type-specific gene expression (Jin et al. 2013a; Lieberman-Aiden et al. 2009a; Schmitt et al. 2016; Nora et al. 2012), as well as misregulation of oncogenes and tumor suppressors in cancer (Taberlay et al. 2016a; Hnisz et al. 2016a; Franke et al. 2016) and other diseases (Li et al. 2012). Identifying changes in chromatin interactions is the next logical step in understanding genomic regulation.

Evolution of Chromatin Conformation Capture (3C) technologies into Hi-C sequencing now allows the detection of “all vs. all” long-distance chromatin interactions across the whole genome (Lieberman-Aiden et al. 2009a; Sanborn et al. 2015). Soon after public Hi-C datasets became available, it was clear that technology- and DNA sequence-driven biases substantially affect chromatin interactions (Yaffe and Tanay 2011a). The technology-specific biases include the cutting length of a restriction enzyme (HindIII, MboI, or NcoI), cross-linking conditions, circularization length, etc. The DNA sequence-driven biases include GC content, mappability, nucleotide composition. Discovery of these biases led to the development of methods for normalizing individual datasets (Cournac et al. 2012; Lieberman-Aiden et al. 2009a; Imakaev et al. 2012a; Yaffe and Tanay 2011a; Knight and Ruiz 2012a). Although normalization of individual datasets improves reproducibility within replicates of Hi-C data (Imakaev et al. 2012a; Yaffe and Tanay 2011a), these methods do not consider biases between multiple Hi-C datasets.

Accounting for the between-dataset biases is critical for the correct identification of chromatin interaction changes between, e.g., disease-normal states, or cell types. Left unchecked, biases can be mistaken for biologically relevant differential interactions. While DNA sequence-driven biases affect two datasets similarly (e.g., CG content of genomic regions tested for interaction differences is the same), technology-driven biases are poorly characterized and affect chromatin interactions unpredictably. Importantly, another source of chromatin interaction differences stems from large-scale genomic rearrangements, such as copy number variations (Servant et al. 2015), a frequent event in cancer genomes (Rickman et al. 2012a). Accounting for such biases is needed for the accurate detection of differential chromatin interactions between Hi-C datasets.

We developed an R package, `HiCcompare2`, for the joint normalization and comparative analysis of multiple Hi-C datasets, summarized as chromatin interaction matrices. Our method is based on the observation that chromatin interactions are highly stable (Dixon et al. 2012; Fudenberg et al. 2016; Rao et al. 2014a; Schmitt et al. 2016), suggesting that the majority of them can serve as a reference to build a rescaling model. We present the novel concept of the MD plot (Minus, or difference vs. Distance plot), a modification of the MA plot (Dudoit et al. 2002a). The MD plot allows for visualizing the differences between interacting chromatin regions in two Hi-C datasets while explicitly accounting for the linear distance between interacting regions. The MD plot concept naturally allows for fitting the local regression model, a procedure termed loess, and jointly normalizing the two datasets by balancing biases between them. The distance-centric view of chromatin interaction differences allows for detecting statistically significant differential chromatin interactions between two Hi-C datasets. We show improved performance of differential chromatin interaction detection when using the jointly vs. individually normalized Hi-C datasets. Our method is broadly applicable to a range of biological problems, such as identifying differential chromatin interactions between tumor and normal cells, immune cell types, and normal tissues/cell types.

2.2 Methods

2.2.1 Visualization of the differences between two Hi-C datasets using an MD plot

A chromosome-specific Hi-C matrix is a square matrix of size $N \times N$, where N is the number of genomic regions of size X on a chromosome. The size X of the genomic regions defines the resolution of the Hi-C data. Each cell in the matrix contains an interaction frequency $IF_{i,j}$, where i and j are the indices of the interacting regions. For this study, data in the sparse upper triangular format from the GM12878 and RWPE1 cell lines were used (Appendix 1 Supplemental Methods).

The first step of the `HiCcompare` procedure is to convert the data into what we refer to as an MD plot. The MD plot is similar to the MA plot (Bland-Altman plot) commonly used to visualize gene expression differences (Dudoit et al. 2002a). M is defined as the log difference between the two data sets $M = log_2(IF_2/IF_1)$, where IF_1 and IF_2 are interaction frequencies of the first and the second Hi-C datasets, respectively. D is defined as the distance between two interacting regions, expressed in unit-length of the X resolution of the Hi-C data. In terms of chromatin interaction matrices, D corresponds to the off-diagonal traces of interaction frequencies (Figure 2.1). Because chromatin interaction matrices are sparse, i.e., contain an excess of zero interaction frequencies, by default only the non-zero pairwise interaction are used for the construction of the MD plot with an option to use partial interactions, i.e., with a zero value in one of the matrices and a non-zero IF in the other.

2.2.2 Joint normalization of multiple Hi-C data using `loess` regression

After the transformation of the data into an MD plot, loess regression (Cleveland 1979) is performed with D as the predictor for M. The fitted values are then used to normalize the original IFs:

$$\begin{cases} log_2(\hat{IF}_{1D}) = log_2(IF_{1D}) + f(D)/2 \\ log_2(\hat{IF}_{2D}) = log_2(IF_{2D}) - f(D)/2 \end{cases}$$

where $f(D)$ is the predicted value from the loess regression at a distance D. The $log_2(\hat{IF})$ values are then anti-logged to obtain the normalized IFs. Note that for both Hi-C datasets the average interaction frequency remains unchanged, as IF_1 is increased by the factor of $f(D)/2$ while IF_2 is decreased by the same amount. Any normalized IFs with values less than one are not considered in further analyses. The joint normalization was tested against five methods for normalizing individual Hi-C matrices, `ChromoR` (Shavit and Lio' 2014a), `ICE` (Imakaev et al. 2012a), `KR` (Knight and Ruiz 2012a), `SCN` (Cournac et al. 2012), `MA` (Lun and Smyth 2015a).

2.2.3 Excluding potentially problematic regions from the joint normalization

The between-dataset biases may occur due to large-scale genomic rearrangements and copy number variants (CNVs), a frequent case in tumor-normal comparisons (Rickman et al. 2012a). Similar to removing other biases, the joint loess normalization removes CNV-driven biases by design, allowing for the detection of chromatin interaction differences within CNV regions. However, CNVs introduce large changes in chromatin interactions (Servant et al. 2015), which may be of interest to consider separately. Therefore, unless cells/tissues with normal karyotypes are compared, we provide functionality for the detection and removal of genomic regions containing CNVs from the joint normalization. The QDNAseq (Scheinin et al. 2014) R package is used to detect and exclude CNVs from the `HiCcompare` analysis. Alternatively, CNV regions can be detected separately and provided to `HiCcompare` as a BED file. Additionally, the `HiCcompare` package includes the ENCODE blacklisted regions for hg19 and hg38 genome assemblies, which can be excluded from further analysis.

2.2.4 Detection of differential chromatin interactions

After joint normalization, the chromatin interaction matrices are ready to be compared for differences. Again, the MD plot is used to represent the differences M between two normalized datasets at a distance D. The jointly normalized M values are centered around 0 and are approximately normally distributed across all distances (Appendix 1 Supplemental Methods). M values can be converted to Z-scores using the standard approach:

$$Z_i = \frac{M_i - \overline{M}}{\sigma_M}$$

where $\overline{M}$ is the mean value of all M's on the chromosome and σ_M is the standard deviation of all M values on the chromosome and i is the ith interacting pair on the chromosome.

2.2.5 Filtering low-abundance interaction pairs

During Z-score conversion, the average expression of each interacting pair is considered. Due to the nature of M, a difference represented by an interacting pair with IFs 1 and 10 is equivalent to an interacting pair of IFs 10 and 100 with both differences producing an M value of 3.32. However, the average expression of these two differences is 5.5 and 55, respectively. Differences with higher average expression are supported by the larger number of sequencing reads and are therefore more trustworthy than the low average expression differences. Thus, we filter out differences with low average expression by setting the Z-scores to 0 when average expression (A) is less than a user set value of A (Appendix 1 Supplemental Methods). Filtering takes place such that the $\overline{M}$ and σ_M are calculated using only the M values remaining after filtering. The Z-scores can then be converted to p-values using the standard normal distribution.

2.2.6 Multiple testing correction

Analyzing Hi-C data for differences necessarily involves testing of multiple hypotheses. Multiple testing correction (False Discovery Rate (FDR) by default) is applied on a per-distance basis by default, with an option to apply it on a chromosomal basis.

2.2.7 Benchmarking the differential chromatin interaction detection

As there is no “gold standard” for differential chromatin interactions, we created such *a priori* known differences by introducing controlled changes to replicate Hi-C datasets (Dozmorov et al. 2010a). To introduce these a priori known differences, we start with two replicates of Hi-C data from the same cell type. It is assumed that any differences in these replicates are due to noise or technical biases. Next, we randomly sample a specified number of entries in the contact matrix. These sampled entries are where the changes will be introduced. The IFs for each of these entries in the two matrices are set to their average value between the replicates, and then one of them is multiplied by a specified fold change. This introduces a true difference at an exact fold change between the two replicates. The benefit of using joint normalization vs. individually normalized datasets was quantified by the improvement in the power of detecting the pre-defined chromatin interaction differences. Standard classifier performance measures (Appendix 1 section 5), summarized in the Matthews Correlation Coefficient (MCC) metric, were assessed. The results of the `HiCcompare` analysis were further compared with those obtained with the `diffHiC` method (Lun and Smyth 2015a).

2.2.8 Example HiCcompare analysis using mouse neuronal differentiation

As an example case for the usage of `HiCcompare`, we performed an analysis to compare the 3D structure of the chromatin between mouse embryonic stem cells (ESC), neural progenitor cells (NPC), and neurons. The data was obtained from a study by Fraser et al. (Fraser 2015) deposited on GEO [GSE59027]. The Hi-C matrices for each cell type were downloaded at 100KB resolution and read into HiCcompare. We performed three comparisons between the cell types, ESC vs. NPC, NPC vs. neuron, and ESC vs. neuron. In each comparison, the data were normalized, low average expression interactions were filtered out, and the differences between the cell types were detected. We also performed a functional enrichment analysis of genes located in differentially interacting regions using KEGG and Gene Ontology analyses using `EnrichR`.

2.2.9 Comparison with `diffHic`

To compare `HiCcompare` with `diffHic`, we performed a `HiCcompare` analysis on RWPE1 Hi-C data (Rickman et al. 2012a) using HiCcompare. This was compared to the analysis performed in the `diffHic` paper (Lun and Smyth 2015a). We performed the analysis at a 1MB resolution as done by Lun et al. Because `diffHic` takes unaligned Hi-C data as input it was not possible to directly compare our method to `diffHic` using introduced known changes. We performed an overlap analysis of the regions detected by our method with the regions detected by `diffHic` (courtesy to Drs. Gordon Smyth and Aaron Lun). Additionally, we compared the fold changes and average expression values of the regions detected by each method.

2.2.10 Comparison with `FIND`

To make a comparison with `FIND` (Djekidel et al. 2018a) we first repeated their comparison of the GM12878 and K562 using `HiCcompare`. Data from GM12878 and K562 were obtained from GEO (GSE63525, samples GSM1551574, GSM1551575, GSM1551620, and GSM1551623) (Rao et al. 2014a). First, the maximum resolution of each dataset was calculated using `Juicer` (Durand et al. 2016a). Next, replicates for each cell type were combined and then input into `HiCcompare` for joint normalization and differential analysis. This was performed at resolutions of 1MB, 100KB, 50KB, 10KB, and 5KB. The differential regions were intersected with the locations of all genes using BEDtools. Then the genes enriched in the differential regions were input into a KEGG pathway enrichment analysis using `EnrichR` (Edward Y Chen and Ma’ayan 2013).

Additionally, the two replicates of GM12878 were used as the basis for comparing `HiCcompare` and `FIND` when *a priori* known differences were introduced into these replicates. For the data to be entered into FIND, we used the VC squared normalization method from `Juicer` as described in the `FIND` paper and the raw data was entered into HiCcompare. We performed this analysis at resolutions of 1MB (we encountered issues due to

extremely long run times of `FIND` when attempting comparisons at higher resolutions) with fold changes of 2, 3, and 5 for the true changes. Standard classifier performance measures were calculated for each method, and MD plots showing where each method was detecting differences were produced.

2.2.11 Concordance between A/B genomic compartments

To assess the effect of normalization of the detection of A/B compartments, they were defined using the Principal Components Analysis (PCA) method (Lieberman-Aiden et al. 2009a) using the raw, jointly and individually normalized RWPE1 Hi-C data (Rickman et al. 2012a). The concordance of compartment detection was evaluated using three metrics: 1) the Pearson correlation coefficient between the vectors of principal components (PCs) detected from raw and normalized data, 2) the overlap of signs of PCs defining A (positive) and B (negative) compartments, and 3) the Jaccard overlap statistics.

2.2.12 Parallelization

The biggest advantage of loess - the ability to model any biases in the data without explicitly specifying them - comes at the cost of increased computation. We implemented a parallelization strategy for processing chromosome-specific chromatin interaction matrices on multiple cores, improving the total run time (Appendix 1 Supplemental Figure 3.1). The parallelization strategy makes use of the Bioconductor `BiocParallel` R package.

2.2.13 Software availability

`HiCcompare` is available on Bioconductor at https://bioconductor.org/packages/HiCcompare/ or on GitHub at https://github.com/dozmorovlab/HiCcompare. The package includes vignettes with test data and documentation for all functions, as well as code to generate all results referenced in this manuscript. `HiCcompare` is released under the MIT open source software license.

2.3 Results

2.3.1 The off-diagonal concept of distance between regions in chromatin interaction matrices

Our study focuses on the joint analysis of multiple Hi-C datasets represented by chromatin interaction matrices, where rows and columns represent genomic regions (bins), and cells contain interaction counts (frequencies). The values on the diagonal trace represent interaction frequencies (IFs) of self-interacting regions. Each off-diagonal trace of values represents interaction frequencies for a pair of regions at a given unit-length distance. The unit-length distance is expressed in terms of resolution of the data (the size of genomic regions, typically measured in millions (thousands) of base pairs, MB (KB)). The concept of considering interaction frequencies at each off-diagonal trace is central for the joint normalization and differential chromatin interaction detection (Figure 2.1).

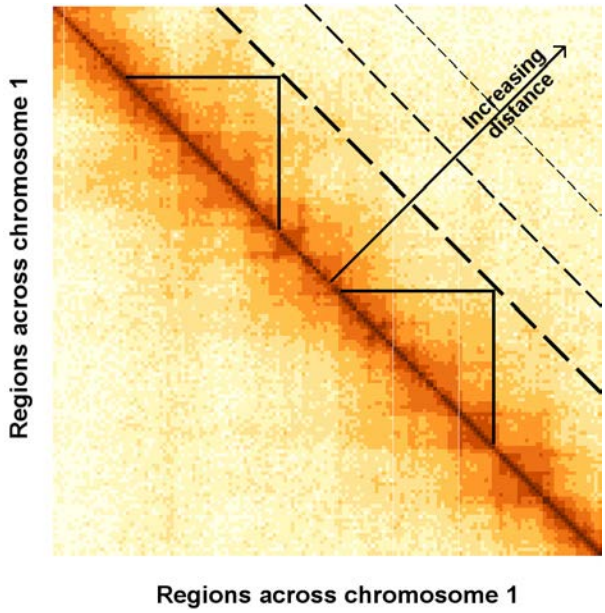

Figure 2.1. Distance-centric (off-diagonal) view of chromatin interaction matrices. Each off-diagonal vector of interaction frequencies represents interactions at a given distance between pairs of regions. Triangles mark pairs of genomic regions interacting at the same distance. Data for chromosome 1, K562 cell line, 50KB resolution, spanning 0 - 7.5Mb is shown.

The interaction frequency drops as the distance between interacting regions increases. Numerous attempts have been made to parametrically model the inverse relationship between chromatin interaction frequency and the distance between interacting regions. However, Hi-C data are affected by technology- and DNA sequence-driven biases (Yaffe and Tanay 2011a; Cournac et al. 2012; Imakaev et al. 2012a), unpredictably altering chromatin interaction frequencies. Consequently, parametric approaches fail to model interaction frequencies across the full range of distances (Sanborn et al. 2015), confirmed by our observations (Figure 1.2).

2.3.2 Elimination of biases in jointly, but not individually, normalized Hi-C data

Discovery of biases in Hi-C data led to the development of numerous methods for normalizing *individual* datasets (Lieberman-Aiden et al. 2009a; Imakaev et al. 2012a; Knight and Ruiz 2012a; Cournac et al. 2012). Although normalization of individual datasets improves reproducibility of replicated Hi-C data (Imakaev et al. 2012a; Yaffe and Tanay 2011a), these methods do not explicitly account for biases between *multiple* Hi-C datasets. The between-dataset biases are particularly problematic when comparing Hi-C datasets between biological conditions (Appendix 1 section 4). When the detection of chromatin interaction differences due to biology, not biases, is important, normalization that removes the between-dataset biases is critical.

To assess the between-dataset biases, we visualize two Hi-C datasets on a single MD plot (see Methods). Briefly, differences in chromatin interaction frequencies (**M**inus) are visualized on a per-unit-length distance basis. Chromatin interactions are highly conserved (Dixon et al. 2012; Fudenberg et al. 2016; Rao et al. 2014a); thus, the majority of the **M** differences should be centered around zero. The MD plot visualization allows us to identify systematic biases appearing as the offset of the cloud of **M** differences from zero. Visualizing replicates of Hi-C data (Gm12878 cell line) showed the presence of biases in the individually normalized datasets (Figure 2.2, Appendix 1 section 4), suggesting that the performance of individual normalization methods may be sub-optimal when comparing multiple Hi-C datasets.

To account for between-dataset biases, we developed a non-parametric joint normalization method that makes no assumptions about the theoretical distribution of the chromatin interaction frequencies. It utilizes the well-

known loess (locally weighted polynomial regression) smoothing algorithm - a regression-based method for fitting simple models to segments of data (Cleveland 1979). The main advantage of loess is that it accounts for any local irregularities *between* the datasets that cannot be modeled by parametric methods. Thus, loess is particularly appealing when normalizing two Hi-C datasets, as the internal biases in Hi-C data are poorly understood (Figure 2.2).

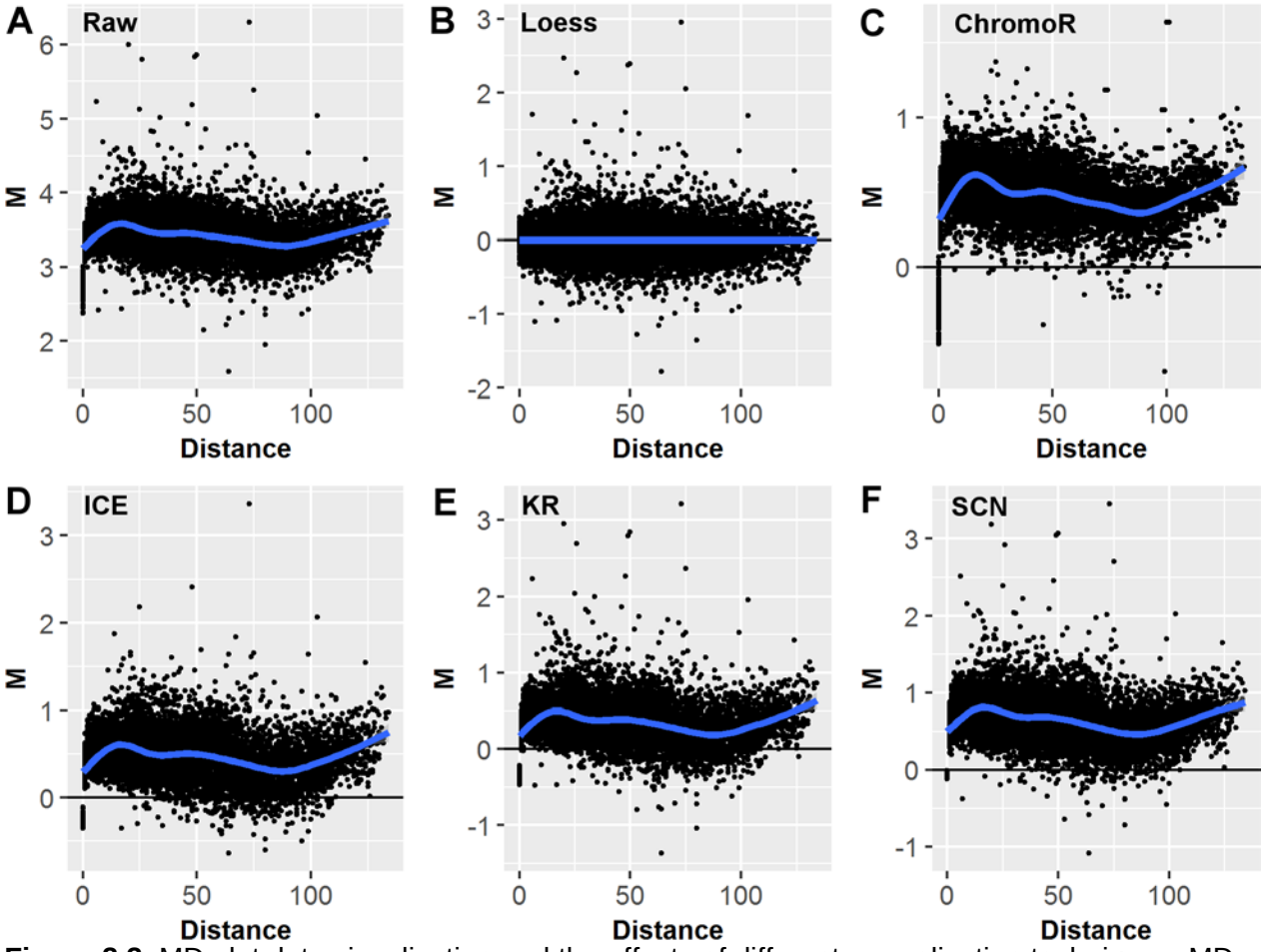

Figure 2.2. MD plot data visualization and the effects of different normalization techniques. MD plots of the differences M between two replicated Hi-C datasets (GM12878 cell line, chromosome 11, 1MB resolution, DpnII and MboI restriction enzymes) plotted vs. distance D between interacting regions. (A) Before normalization, (B) after loess joint normalization, (C) `ChromoR`, (D) Iterative Correction and Eigenvector decomposition (`ICE`), (E) Knight-Ruiz (`KR`), (F) Sequential Component Normalization (`SCN`).

Existing Hi-C data at high-resolutions (e.g., 10 kb) still suffer from a limited dynamic range of chromatin interaction frequencies, with the majority of them being small or zero, especially at large distances between interacting regions. This sparsity places limits on loess joint normalization, as it builds a rescaling model from many non-zero pairwise comparisons. A way to alleviate this limitation is to consider interactions only within a range of short interaction distances, where genomic regions interact more frequently, and the proportion of zero interaction frequencies is the lowest. Our evaluation of loess joint normalization showed it performs best at resolutions between 1MB and 50KB (Appendix 1 section 4, Appendix 1 section 7). The issue of sparsity limiting the usefulness of loess normalization will be alleviated as sequencing techniques continue to improve and Hi-C datasets with deeper sequencing become available.

2.3.3 Detecting differential chromatin interactions

To benchmark our detection of significant chromatin interaction differences, we introduced *a priori* known chromatin interaction differences in replicate data from the GM12878 cell line. The benefits of the joint

normalization vs. individually normalized datasets were evaluated in detecting the known differences. HiCcompare can detect most of the added differences with a relatively low number of false positives across the range of fold changes (Table 2.1, Appendix 1 section 5).

Table 2.1. Evaluation of the effect of normalization on differential chromatin interaction detection. Matthews Correlation Coefficient of detecting 200 controlled differences in jointly (HiCcompare) vs. individually normalized Gm12878 datasets, chromosome 1, 1KB resolution. Matrices were normalized with methods corresponding to column labels; differences were detected using HiCcompare.

Fold change	HiCcompare	MA	ICE	SCN	KR	ChromoR
2	0.847	0.823	0.835	0.768	0.748	0.149
3	0.973	0.934	0.802	0.721	0.764	0.380
4	0.995	0.98	0.953	0.881	0.868	0.532

2.3.4 Example HiCcompare analysis using mouse neuronal differentiation

As expected, the ESC vs. neuron had the largest number of differentially interacting regions at 951 (FDR < 0.05). The ESC and NPC had 279 differentially interacting regions, and the NPC and neuron had only 127 differentially interacting regions. These differences expectedly suggest that the undifferentiated ESCs and fully differentiated neurons have many chromatin interaction differences, while the intermediate neural progenitor cells have fewer differences when compared with either ESCs or neurons. These observations suggest that the chromatin structure plays a key role in the process of cell differentiation.

The enrichment analysis for the ESC vs. the neuron found genes enriched in protein binding function, ion channel regulator activity, and “Axon guidance” pathway among others (Appendix 2). The enrichment of these pathways outlines the ESC-to-neuron differentiation processes that are governed by changes in the 3D structure of the genome. When comparing the ESC and NPC cells, genes were found to be enriched in voltage-gated calcium channel activity, ion transporters, and serotonin metabolic processes (Appendix 3). The enrichment results between the NPC and neuron had fewer results but included IgG receptor activity and binding and cytoskeletal protein binding (Appendix 4). These results indicate that the changes in the chromatin structure contain functionally relevant genes for the cell differentiation process.

The results of this HiCcompare analysis show that our methods are capable of detecting biologically meaningful differences in chromatin conformation when comparing different cell types. Together with the results of Fraser et al. (Fraser 2015), the HiCcompare results indicate that the cellular differentiation process involves structural changes of the chromatin, likely leading to the changes in gene expression and the associated biological pathways.

2.3.5 Comparison with diffHiC

The diffHiC pipeline was designed to process raw Hi-C sequencing datasets and detect chromatin interaction differences using the generalized linear model framework developed in the edgeR package (Lun and Smyth 2015a). We compared the results of Hi-C data analyzed in the diffHiC paper (human prostate epithelial cells RWPE1 over-expressing the EGR protein or GFP (Rickman et al. 2012a)) with the results obtained by HiCcompare.

To compare HiCcompare with diffHic, we performed a HiCcompare analysis on the RWPE1 Hi-C data (Rickman et al. 2012a). This was compared to the analysis performed in the diffHic paper (Lun and Smyth 2015a). We performed the analysis at a 1MB resolution as described in the diffHic paper. diffHic detected a total of 5,737 significant differences (FDR < 0.05), while HiCcompare tended to be more conservative, detecting 680 differences (FDR < 0.05) and 5,215 differences when multiple testing correction was not applied (p-value < 0.05). Of the 680 differences, 208 overlapped with the regions detected by diffHic. Surprisingly, although diffHiC used CNV correction in their analysis, 2,567 (44.7%) of the detected differentially interacting regions overlapped with CNV regions detected in our analysis or blacklisted regions. diffHic tended to detect differentially interacting regions with smaller fold changes as compared to HiCcompare, and at shorter distances between interacting regions, while HiCcompare can detect differences across the full range of

distances (Section 6, Additional File 1). These results suggest that detecting chromatin interaction differences represented in the MD coordinates, as implemented in HiCcompare, may be useful in detecting large chromatin interaction differences across the full range of distances, potentially having a more significant biological effect.

2.3.6 Comparison with FIND

The recently published FIND tool uses a spatial Poisson process to detect differences between two Hi-C experimental conditions (Djekidel et al. 2018a). FIND is presented as a tool for high-resolution Hi-C data and treats interactions as spatially dependent on surrounding interactions. To compare HiCcompare with FIND, we performed a comparative analysis between Hi-C data from K562 and GM12878 cells lines (Appendix 1 section 7) as done in the FIND paper (Djekidel et al. 2018a). The maximum resolution of each Hi-C matrix was calculated using the calculate_map_resolution.sh function from Juicer (Durand et al. 2016a). Briefly, two replicates for each cell line were obtained (see Methods), and the replicate contact matrices were combined for the HiCcompare analysis. HiCcompare was used to jointly normalize the data between the cell lines and then detect differences. HiCcompare analyses were performed at 1MB, 100KB, 50KB, 10KB, and 5KB resolutions. Additionally, the analyses of GM12878 and K562 were used to compare the run times of HiCcompare and FIND (Appendix 1 section 7).

The number of differences detected by HiCcompare at 5KB resolution was much lower than the number FIND detected (~150,000) (Djekidel et al. 2018a). The drop off of the number of differential interactions detected at high-resolution by HiCcompare can be explained by the sparsity and the limited dynamic range of interaction frequencies at 5KB resolution. Additionally, the large number of differences detected by FIND at 5KB resolution is questionable given that the maximum resolution of the K562 and GM12878 data was found to be ~39KB and ~9KB, respectively (Appendix 1 section 7).

The differentially interacting regions detect by HiCcompare at different resolutions were intersected with gene locations, and a KEGG pathway enrichment analysis was performed. The enrichment analysis showed that many of the differential regions contained genes involved in the immune system (Table 2.2). We also found that the enrichment analyses of HiCcompare-detected differences at each resolution were relatively consistent further indicating the strength of HiCcompare at detecting biologically relevant differences across data resolutions. Despite the differences in resolution of data used for differential analysis (5 kb for FIND and 50 kb - 1 Mb for HiCcompare) the enrichment analysis of HiCcompare-detected differences identified pathways related to the immune system, similar to the results of the FIND analysis. These observations suggest that both methods can detect biologically significant differences.

Table 2.2. Gene enrichment results for HiCcompare analyses. KEGG pathways and their corresponding FDR-corrected p-values for the enrichment analyses of HiCcompare-detected differences at 1MB, 100KB, and 50KB resolutions. Differentially interacting regions detected by HiCcompare were intersected with gene locations, and the overlapping genes were tested for enrichment using EnrichR (Edward Y Chen and Ma'ayan 2013).

Pathway	1MB	100KB	50KB
Systemic lupus erythematosus	3.807e-06	6.302e-17	1.025e-02
Antigen processing and presentation	3.807e-06	6.808e-01	9.974e-01
Staphylococcus aureus infection	8.170e-03	2.354e-01	7.604e-01
Viral myocarditis	8.170e-03	1.038e-01	9.657e-01
Allograft rejection	8.170e-03	1.518e-01	9.974e-01
Viral carcinogenesis	3.327e-02	3.659e-08	3.273e-01
Pathways in cancer	9.162e-01	2.236e-02	9.409e-01

To compare the performance of FIND and HiCcompare we used replicated data for GM12878 cells. The GM12878 replicates are expected to contain minimal differences, thus suitable for introducing *a priori*

controlled changes (see Methods) and applying both tools to detect these changes. HiCcompare successfully detected the majority of the controlled changes while FIND detected smaller differences and was missing most of the introduced controlled changes (Appendix 1 section 7). Additionally, we found that the run time of FIND on Hi-C matrices at resolutions between 100KB and 10KB was extremely long (>72 hours) even when run in parallel on 16 cores, while HiCcompare was able to complete an analysis within minutes (Appendix 1 Supplemental Figure 3.1). These results further strengthen the notion that HiCcompare detects large chromatin interaction differences potentially having a larger biological impact on genome structure, and does it across the full range of distances.

2.3.7 Preservation of A/B compartments

A/B compartments are the best known genomic structures that can be detected from Hi-C data (Lieberman-Aiden et al. 2009a). A/B compartments are large genomic features which can be detected in Hi-C contact matrices. A compartments correspond to open chromatin and gene expression and B compartments are associated with closed chromatin and low gene expression. To understand the consequences of the joint vs. individual normalization methods on the detection of A/B compartments we compared principal components defining compartments in raw vs. normalized data. A/B compartments detected following joint normalization were the most similar to those detected in the raw data (Table 2.3). These results suggest that the joint HiCcompare normalization preserves properties of Hi-C data needed for the accurate detection of A/B compartments.

Table 2.3. Similarity between A/B compartments detected following various normalization methods. "Correlation" - Pearson correlation coefficient between principal components defining A/B compartments in raw vs. normalized Hi-C data; "Prop. Match Sign" - the proportion of regions with matching signs defining A/B compartments; "Jaccard A/B" - Jaccard overlap statistics between A/B compartments, respectively. All values represent averages over all chromosomes.

Comparison	Mean Absolute Correlation	Mean Percentage	Jaccard A	Jaccard B
Loess vs. Raw	0.9954	0.8537	0.7971	0.7823
MA vs. Raw	0.9950	0.8539	0.7881	0.7706
ICE vs. Raw	0.9795	0.7850	0.6731	0.6277
KR vs. Raw	0.9489	0.7771	0.5945	0.5000
SCN vs. Raw	0.9309	0.8083	0.6134	0.5495
ChromoR vs. Raw	0.8093	0.6810	0.5210	0.4803

Taken together, our results demonstrate the importance of joint normalization when comparing Hi-C datasets. We introduce the concept of a distance-centric view of Hi-C data, implemented as an intuitive distance-centric visualization of two Hi-C datasets on the MD plot. The MD plot representation allows for joint loess normalization that improves power in detecting true chromatin interaction differences and preserves data properties needed for the accurate detection of A/B compartments. The HiCcompare R package implements the visualization, joint normalization, and differential chromatin interaction detection algorithms, allowing for the comparison of Hi-C data.

2.4 Discussion

This work introduces three novel concepts for the joint normalization and differential analysis of Hi-C data, implemented in the HiCcompare R package. First, we introduce the representation of the differences between two Hi-C datasets on an MD plot, a modification of the MA plot (Dudoit et al. 2002a). Importantly, we consider the data on a per-distance basis, allowing the data-driven normalization of global biases without distorting the relative distribution of interaction frequencies of the interacting regions. Second, we implement a non-parametric loess normalization method that minimizes bias-driven differences between the datasets. There is compelling evidence that non-parametric normalization methods, such as quantile- and loess normalization, are particularly suitable for removing between-dataset biases (Shao et al. 2012; Bolstad et al. 2003), confirmed

by our application of loess to the joint normalization of Hi-C data. Third, we develop and benchmark a simple but rigorous statistical method for the differential analysis of Hi-C datasets.

The importance of joint normalization in removing between-dataset biases has been demonstrated using the MA normalization introduced in the `diffHiC` R package (Lun and Smyth 2015a). `MA` normalization uses a similar concept of representing measures from two datasets on a single plot (Lun and Smyth 2015a), except it uses the **A**verage chromatin interaction frequency as the X-axis instead of the **D**istance. MA normalization performed second to `HiCcompare` (Table 2.1, Appendix 1 section 5). This may be due to the power-law decay of interaction measures leading to the limited dynamic range of average chromatin interaction frequencies and making fitting a loess curve difficult. Instead, the more balanced representation of chromatin interaction differences **M** (Y-axis) as a function of distance **D** (X-axis) improves the performance of loess fit for the joint normalization and the subsequent detection of chromatin interaction differences.

The discrepancy of differential chromatin interaction detection between `diffHiC` and `HiCcompare` (Appendix 1 section 6) could arise from multiple factors. `diffHiC`'s implementation of MA normalization favors differences at shorter distances and small fold changes while HiCcompare's loess fitting through the MD plot allows for the detection of large chromatin interaction differences across the full range of interaction frequencies (Appendix 1 section 6). `diffHiC` operates on logCPM counts while `HiCcompare` uses log interaction frequency counts. `diffHiC` uses enzyme cut sites to define bins when partitioning the genome while `HiCcompare` uses fixed bin sizes. `diffHiC` uses median inter-chromosomal interaction frequency to filter low-abundance bin pairs while `HiCcompare` filters based on average IFs of the chromosome being considered. Finally, the RWPE1 data analyzed by `diffHiC` is relatively sparse even at 1MB resolution, potentially interfering with HiCcompare's statistical analyses. In summary, `diffHiC` and `HiCcompare` may provide complementary views on chromatin interaction differences, with `HiCcompare` being better suited for removing the between-datasets biases and the detection of biology-driven chromatin interaction differences.

In our comparison with `FIND` (Appendix 1 section 7), we found that `HiCcompare` performed better than `FIND` on data at resolutions between 1MB and 10KB. As most publicly available Hi-C data is too sparse to make meaningful inferences at resolutions greater than this, `HiCcompare` looks to be the better choice for detecting differences on most currently available data. In the case of extremely high-resolution Hi-C data, `FIND` may be able to pull out more significant differences between two experimental conditions albeit at the expense of significantly longer run times. Comparing our gene enrichment results for GM12878 vs. K562 with those presented in (Djekidel et al. 2018a), both methods were able to detect differences in regions involved in the immune system as would be expected to occur for these cell types.

Despite the ability of Hi-C technology to simultaneously capture all genomic interactions, current resolution of Hi-C data (1MB - 1KB) remains insufficient to resolve individual *cis*-regulatory elements (~100b-1KB). Alternative techniques, such as ChiA-PET (Fullwood et al. 2009), Capture Hi-C (Mifsud et al. 2015) have been designed to identify targeted 3D interactions, e.g., between promoters and distant regions. These data require specialized normalization (Cairns et al. 2016) and differential analysis (Lareau and Aryee 2017) methods. Our future goals include extending the loess joint normalization method for chromosome conformation capture data other than Hi-C.

Chapter 3: Aim 2 - Application of `HiCcompare` to human brain data

3.1 Introduction

The human brain is a complex organ composed of distinct anatomical and functional regions and cell types. Being the central organ for human cognition, it is also one of the most difficult organs to study (Birdsill et al. 2011; Popova et al. 2008). While gene expression and epigenomic changes across human brain regions and cell types have been characterized (Kang et al. 2011; Strand et al. 2007; GTEx Consortium 2013), the dynamics of 3D chromatin interactions integrated with "omics" changes remain undefined. It is of interest to determine if there are 3D structural differences between distinct regions of the brain. Differences in chromatin structure could be integral to the tissue differentiation observed between the brain regions. These structural changes in the chromatin found between regions of the brain can then be associated with differences in gene expression or enrichment of other epigenomic features using standard enrichment techniques. To study the

links between 3D structure and tissue differentiation we have collected Hi-C data for the amygdala and the prefrontal cortex (PFC) of a single human fetal brain.

The amygdala is a distinct anatomical region of the brain located deep within the temporal lobes. It is associated with memory and emotional responses. The amygdala is part of the limbic system, which plays a role in several cognitive functions including long-term memory, emotion, and olfaction. This brain region is also implicated in fear responses, which are linked to the expression of certain signaling pathways. The amygdala is an older structure of the brain on the evolutionary scale and is present in most complex vertebrates.

The prefrontal cortex is the region of the cerebral cortex in the frontal lobe. This region of the brain is involved in many higher functions of cognition. It is thought to be involved in planning complex behaviors and decision making. The prefrontal cortex is involved in top-down processing of behaviors, which are the product of sensory input or thoughts that are subject to rapid changes (Miller EK 2001). The prefrontal cortex is highly interconnected, with links to most other regions of the brain. The classic case showing the functions of the prefrontal cortex is that of Phineas Gage. Gage had an iron rod driven through his left frontal cortex in an accident but survived. His personality changed after the accident yet his memory and motor functions were not affected.

A previous study by Won et al. (Won et al. 2016) performed a Hi-C analysis on samples from the developing human brain. Briefly, they took samples from the cerebral cortex at the peak of neurogenesis from the subcortical plate and the germinal zone. They compared the contact profiles and found them to be similar between replicates and individuals. Additionally, they observed switching between type A and type B compartments (type A compartments are associated with open chromatin and gene expression, type B compartments with closed chromatin and low expression). They found that regions associated with promoters and enhancers were more likely to be interacting in 3D space. However, for their comparisons, they used `ICE` normalization. `ICE` is not a joint normalization technique, which could mean the data still contained inter-dataset biases. Furthermore, their identification of interacting regions was performed only considering a single Hi-C matrix at a time. They declared significance for a given Hi-C contact based on the probability of observing a stronger contact under a fitted Weibull distribution matched by chromosome and distance (Won et al. 2016). This is a widely used approach for finding significant contacts within a Hi-C matrix; however, it does not give a statistical measure of the differences in IFs between Hi-C datasets.

Here we propose to use our methods developed in Aim 1, `HiCcompare`, to compare the amygdala and the prefrontal cortex Hi-C data from a human fetal brain. Joint normalization as implemented in HiCcompare will help to remove between dataset biases and aid in comparison. Additionally, we will be able to directly look for differences in the Hi-C contacts of the amygdala and the prefrontal cortex. Finding such statistical differences as opposed to merely overlapping regions determined as significant from a single Hi-C matrix will allow for a better comparison of the regions of the human brain.

3.2 Methods

3.2.1 Generation of Hi-C libraries

Samples were obtained from the amygdala and the dorsolateral prefrontal cortex of a single human fetal brain. Two Hi-C libraries were produced using the samples, one from the amygdala and one from the prefrontal cortex (PFC). The initial replicate from the PFC was not sequenced deep enough, so an additional replicate was produced. The Sau3AI restriction enzyme (cuts at the GC-balanced site “GATC”) was used for generating the Hi-C libraries. Quality of the data (overrepresented sequences, GC content, percent deduplicated, total reads) were assessed using FastQC (Andrews 2010).

3.2.2 Processing of raw Hi-C data

The raw Hi-C data was processed using the `HiC-Pro` pipeline (Servant et al. 2015). The `HiC-Pro` pipeline was used to align the raw data to the hg19 reference genome, perform quality control measures, and produce contact matrices. The built-in ICE normalization procedures of HiC-Pro were not used as we will use HiCcompare’s normalization functionality instead. Only intrachromosomal contact matrices were used for this study. Contact matrices were produced in resolutions of 1MB, 500KB, 100KB, 50KB, 20KB, and 10KB.

3.2.3 Determining the maximum resolution of the data

The maximum usable resolution for each dataset was calculated using our modified version of the algorithm described for calculating map resolution in (Rao et al. 2014a). Briefly, to determine the maximum resolution, we use a very high-resolution dataset, i.e., 100bp. A vector composed of the number of reads for every region in the genome is created. This is equivalent to taking the row (or column) sums of the genome-wide Hi-C matrix.

Next, the percentage of regions with a count greater than 1,000 was calculated. If this percentage is less than 80%, the bins were combined and their counts were summed to bump the resolution of the matrix up to the next step. The number of new bins and the number of current bins being combined is calculated as follows.

$$N_{new\ bins} = ceiling(G/R)$$

Where G is the total genome length and R is the new resolution determined by the original resolution + the step size.

$$N_{per\ bin} = ceiling(N_{bins}/N_{new\ bins})$$

Where N_{bins} is the previous number of bins. These new bins are then assigned to the original bins (if $N_{new\ bins}$ is not a multiple of N_{bins}, the final bin will contain a number of combined bins less than $N_{per\ bin}$). The counts for the new bins are calculated simply by summing all counts of the original bins being combined. Then once again the percent of bins with a count greater than 1,000 is calculated. This process of combining bins and increasing the resolution is repeated until the percent of bins with counts greater than 1,000 is 80% or more. The resolution once this condition is met is the minimum resolution that the Hi-C dataset can support. The maximum resolutions for the amygdala dataset was 17KB, the first replicate of the PFC data was 33KB, and the second replicate of the PFC data was 17KB.

3.2.4 CNV Analysis

CNVs can potentially change the structure of the DNA, and thus it is important to map where they occur. In some cases, it may be necessary for CNV regions to be filtered out from the Hi-C data as they could potentially result in false positives. The wrapper function included in `HiCcompare` for the `QDNAseq` R package was used to detect copy number variations (CNVs) (Scheinin et al. 2014) in our Hi-C data. CNVs were checked for in all the datasets to be analyzed at resolutions of 1MB, 100KB, and 10KB. The CNV regions were compared against regions known to contain repetitive sequences. CNV regions were not excluded from the `HiCcompare` analysis.

3.2.5 `HiCcompare` analysis

`HiCcompare` was used to jointly normalize the datasets and perform three comparisons, amygdala vs. PFC replicate 1, amygdala vs. PFC replicate 2, and PFC 1 vs. PFC 2 at resolutions of 1MB, 100KB, and 20KB. Briefly, `HiCcompare`'s loess normalization was performed on the MD plots for each of the three comparisons. Filtering based on the Average (A) expression values of the interaction frequencies was performed. For the 1MB and 20KB resolution comparisons, the bottom 10th percentile of A values were filtered out. For the 100KB comparisons, interactions with $A < 8$ were filtered out. To determine the filtering level, we first copy the real Hi-C matrix then add in random noise to simulate a technical replicate. Then a specified number of differences at a controlled fold change are added into the fuzzed matrix, and the `HiCcompare` difference detection algorithm is used to check for the true differences added to the matrix. This process is repeated over a range of minimum A values so that an optimal filtering level can be determined where the Matthews Correlation Coefficient (MCC) is maximized without filtering too much of the data. After filtering, the normalized M values were then converted to Z-scores and compared to the standard normal distribution. False discovery rate (FDR) multiple testing correction was applied on a per-distance basis to the p-values. See section 2.2 for a full description of the methods used in HiCcompare.

3.2.6 Enrichment analysis

The regions detected as differentially interacting by the `HiCcompare` analysis were exported as `.BED` files and intersected with a list of the regions for all human genes using BEDtools. This was performed separately for each the 1MB, 100KB, and 20KB analyses. The resulting genes enriched in the differential regions were entered into an enrichment analysis. First, these genes were tested for enrichment in KEGG pathways using the `enrichR` R package. The top most significant pathways were reported. Next, we produced enrichment dot plots using the `clusterProfiler` R package for the most significant KEGG pathways. The pathways found to be enriched in the significant regions were assessed for their fit to the data.

3.2.7 Regions detected multiple times

We used Manhattan plots to identify "hotspot" regions which interact significantly multiple times. The genome was broken down into a list of all linear bins at each analyzed resolution (1MB, 100KB, 20KB) and the number of times each bin was involved in a significant differential interaction was counted. This was performed for each of the three comparisons. The regions detected ≥ 2 times were determined to be hotspot regions. These regions were further investigated for enrichment of genomic and epigenomic features. The genes enriched in the hotspot regions were overlapped in Venn diagrams to look for similar genes appearing between the three comparisons.

3.2.8 RNA-seq on regions of the brain

As we were not able to obtain RNA-seq data directly from the same samples used to produce the Hi-C data, we downloaded RNA-seq data from a total of 229 samples from the human brain that were publicly available on the GTEx Portal. The GTEx 2016 dataset contained 100 samples from the amygdala and 129 samples from the frontal cortex. We input this RNA-seq data into edgeR to perform a differential expression analysis. Reads were filtered and normalized, and the Quasi-likelihood F-test was performed to test for differential expression between the brain regions. The top 1,000 DE genes (FDR < 0.0001) were input into a KEGG enrichment analysis to check for enrichment in pathways.

3.2.9 TAD analysis

Topologically Associated Domains (TADs) are organized units of chromatin structure that are characterized by groups of genomic loci with high levels interacting within the group and low levels of interaction outside the group (Dixon et al. 2012; Kellen G Cresswell 2019). We used SpectralTAD (Kellen G Cresswell 2019) to detect hierarchical TADs on the amygdala and PFC Hi-C data. We then calculated the Jaccard overlap between the TAD boundaries of the two datasets to determine the amount of overlap in TADs between the two brain regions. To determine if the differential regions detected by HiCcompare are occurring within TAD boundaries we checked for their enrichment within the TAD boundaries using the permutation test functionality provided by multiHiCcompare (John C Stansfield 2019).

3.3 Results

3.3.1 Exploratory analysis

Shown in Table 3.1, the quality of sequencing was generally good; however, the first round of sequencing of the PFC had a lower number of reads compared to the amygdala. Due to the poor sequencing quality of the PFC, the library was sequenced again at a deeper level.

Table 3.1. Quality Control for Brain Hi-C Data. Only the Amygdala and the second replicate of the PFC data are displayed. R1 and R2 indicate the number of the file for the pair end fastq files input into FastQC.

File	Basic Statistics	Overrepresented sequences	GC Content	Deduplicated	Total Reads
Amygdala R1	✓	✓	✓	87.60%	414,846,218
Amygdala R2	✓	✓	✓	88.98%	414,846,218
PFC R1	✓	!	✓	81.89%	612,633,643
PFC R2	✓	✓	✓	84.65%	612,633,643

The maximum resolution of the Hi-C data for each dataset was obtained. The amygdala and second cortex replicate were determined to have a maximum resolution of 17KB while the first PFC replicate was determined to have a maximum resolution of 33KB. Thus, 20KB was chosen as the maximum resolution to be used for the `HiCcompare` analyses. The results of the comparison between the amygdala and PFC 1 should not be considered as strongly as the results from the comparison between the amygdala and PFC 2. For the 1MB data, the proportion of interactions with an IF = 0 was calculated in addition to the total number of reads. These results are displayed in Table 3.2.

Table 3.2. Proportion of 0 and total number of reads by Hi-C dataset at 1MB resolution. The cortex combined file represents the combination of the two cortex replicates into a single contact map.

	Proportion 0	Number of reads
PFC 1	0.05865	70572018
PFC 2	0.05624	136825141
PFC Combined	0.05617	205336613
Amygdala	0.05583	131329303

3.3.2 CNV regions

At 1MB resolution, 16 CNVs were detected which occurred on chromosomes 9 and 19. At 100KB resolution, a total of 367 CNVs were detected. At 10KB resolution, a total of 5,029 CNVs were detected. The distribution of the CNVs at 100KB and 10KB resolution are displayed in Table 3.3. The CNV regions detected were not filtered out for the `HiCcompare` analysis.

Table 3.3. Distribution of CNVs at 100KB and 10KB resolution.

resolution	chr	CNV Freq
100KB	chr1	130
100KB	chr2	35
100KB	chr5	17
100KB	chr9	139
100KB	chr10	31
100KB	chr11	15
10KB	chr1	812
10KB	chr2	558
10KB	chr3	21
10KB	chr4	31
10KB	chr5	273
10KB	chr6	34
10KB	chr7	358
10KB	chr8	186
10KB	chr9	1118
10KB	chr10	334
10KB	chr11	48
10KB	chr12	3
10KB	chr14	106
10KB	chr15	355
10KB	chr16	319
10KB	chr17	141
10KB	chr19	28
10KB	chr20	72
10KB	chr21	69
10KB	chr22	163

3.3.3 Regions differing between the amygdala and prefrontal cortex

HiCcompare was used to jointly normalize the datasets for the following comparisons: amygdala vs. PFC 1, amygdala vs. PFC 2, and PFC 1 vs. PFC 2. These comparisons were performed at 1MB, 100KB, and 20KB resolution. After joint normalization, differences were detected between the datasets. The differences detected for each comparison are listed in Tables 3.4-3.6. Except for the 1MB resolution comparison, more differences were found between the amygdala and PFC than between the two PFC replicates. A similar number of differences were found between the amygdala vs. PFC 1 and amygdala vs. PFC 2 comparisons. At 20KB resolution, there were very few differences found between the amygdala vs. PFC 1 and PFC 1 vs. PFC 2 comparisons compared to the amygdala vs. PFC 2 comparison. This is likely due to the fact that the maximum resolution of PFC 1 was 33KB and thus the PFC 1 dataset was too sparse to find many meaningful differences.

Table 3.4. Number of differences detect by `HiCcompare` for each chromosome by each comparison at 1MB resolution.

Chr	amygdala vs PFC 1	amygdala vs PFC 2	PFC 1 vs PFC 2
chr1	74	59	217
chr2	49	59	88
chr3	19	7	91
chr4	41	29	83
chr5	25	11	77
chr6	53	31	93
chr7	27	21	47
chr8	25	19	58
chr9	73	53	121
chr10	7	7	47
chr11	23	22	42
chr12	17	21	58
chr13	8	10	24
chr14	9	8	34
chr15	10	4	23
chr16	20	16	33
chr17	11	11	25
chr18	11	18	19
chr19	4	7	17
chr20	19	23	14
chr21	6	2	23
chr22	3	2	8
chrX	30	20	63

Table 3.5. Number of differences detect by `HiCcompare` for each chromosome by each comparison at 100KB resolution.

Chr	amygdala vs PFC 1	amygdala vs PFC 2	PFC 1 vs PFC 2
chr1	368	336	170
chr2	401	394	232
chr3	342	344	218
chr4	447	344	254
chr5	332	278	183
chr6	288	314	214
chr7	261	221	184
chr8	248	254	162
chr9	197	174	160
chr10	192	175	137
chr11	212	182	146
chr12	209	203	149

chr13	228	186	139
chr14	146	156	113
chr15	112	130	75
chr16	135	93	68
chr17	85	66	77
chr18	193	148	117
chr19	82	89	45
chr20	121	73	67
chr21	84	47	42
chr22	48	56	38
chrX	170	163	73

Table 3.6. Number of differences detect by `HiCcompare` for each chromosome by each comparison at 20KB resolution.

Chr	amygdala vs PFC 1	amygdala vs PFC 2	PFC 1 vs PFC 2
chr1	1	234	5
chr2	1	19	1
chr3	3	0	1
chr4	2	177	3
chr5	1	82	8
chr6	0	156	6
chr7	0	26	4
chr8	0	122	2
chr9	2	6	2
chr10	0	128	5
chr11	0	33	3
chr12	0	45	4
chr13	0	51	7
chr14	0	33	6
chr15	1	57	1
chr16	8	4	5
chr17	0	2	2
chr18	0	72	7
chr19	4	8	0
chr20	1	2	2
chr21	0	0	5
chr22	1	0	0
chrX	1	164	12

We also checked for differentially interacting regions that were highly active. The number of times each region in the genome was found to be involved in a significant differential interaction was counted. Some regions were detected as many as eight times while the majority of regions were detected zero or one times. We set detection of 2 or more times as the criteria for a region to be called a highly differentially interacting region.

Figure 3.1 displays a Manhattan plot of the number of times each region was detected on chromosome 1. The highly differentially interacting regions were then used for the enrichment analysis.

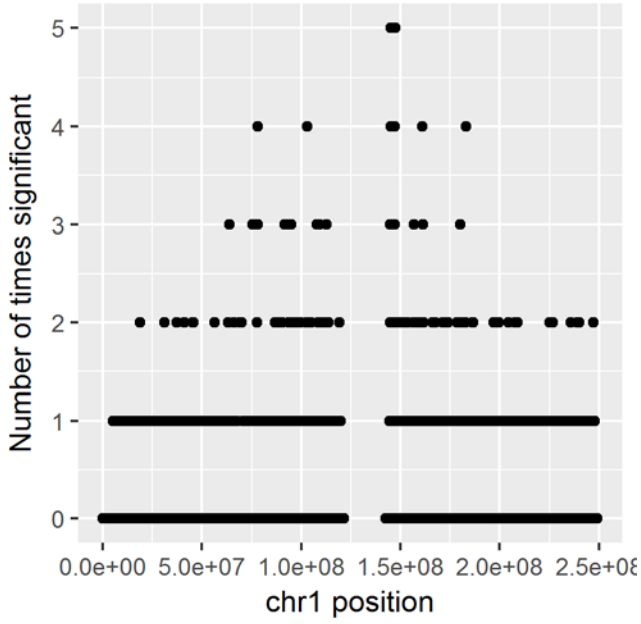

Figure 3.1. Manhattan plot for the number of times each region on Chromosome 1 was detected as significantly differentially interacting in the amygdala vs. PFC 2 comparison at 100KB resolution.

3.3.4 Genes located in highly interacting regions show relevant pathways

We listed all genes whose genomic coordinates intersected with the regions detected two or more times by HiCcompare. We then performed an enrichment analysis on these genes. The top pathways from the enrichment analysis are listed in Table 3.7. Many of the pathways detected by the enrichment analysis are signaling pathways. This indicates that differences in the 3D conformation of the DNA could influence cell signaling and the cellular differentiation processes. We found the “Axon guidance”, “Oxytocin signaling”, and “regulation of actin cytoskeleton” pathways were enriched in these differential regions. This further suggests that differences in the 3D DNA structure could be related to the differences observed between regions of the human brain. The differential regions enriched for the oxytocin signaling pathway may represent one of the ways DNA structure influences the differences of function between the amygdala and PFC. Oxytocin signaling in the amygdala has been found to be required for male mice to have a sexual preference for female mice (Yao 2017). Additionally, oxytocin signaling in the amygdala of rats was linked to the expression of context-conditioned fear responses (Emma J. Campbell-Smith and Westbrook 2015). However, in humans diagnosis of major depressive disorder and bipolar disorder has been associated with high mRNA expression of oxytocin receptors in the PFC compared to control subjects (Lee et al. 2018). These other studies suggest that oxytocin signaling is normal in the amygdala, but high expression of oxytocin receptors in the PFC could be related to disease. This explains why DNA regions associated with oxytocin signaling may have different structural conformations in the amygdala compared to the PFC.

Table 3.7 KEGG pathway enrichment results using the genes contained in the regions detected as differentially interacting.

Pathway	P-value	Q-value
TNF signaling	2.954e-06	2.112e-04
Regulation of actin cytoskeleton	6.753e-06	3.863e-04
MAPK signaling pathway	1.769e-05	7.978e-04
Focal adhesion	2.791e-05	8.261e-04
Adrenergic signaling in cardiomyocytes	7.185e-05	1.514e-03
Rap1 signaling pathway	8.948e-05	1.599e-03
Adherens,junction	1.466e-04	2.466e-03
Oxytocin signaling pathway	6.196e-04	6.563e-03
NF-kappa B signaling pathway	3.114e-04	3.872e-03
Axon guidance	9.145e-04	9.341e-03

3.3.5 Differential genes share similar pathways as Hi-C data

We performed a differential expression analysis between samples from the amygdala and the frontal cortex using data acquired from GTEx Portal. The differential genes were detected and the top 1,000 genes with an FDR < 0.0001 were input into a KEGG enrichment analysis. Several pathways detected using the genes enriched in differential regions detected by HiCcompare were also found to be enriched by the differential genes detected in the RNA-seq analysis (Tables 3.7 & 3.8). Many of the pathways enriched by the DE genes are relevant to brain function such as axon guidance, Dopaminergic synapse, and oxytocin signaling along with several other signaling pathways (Table 3.8). The overlap we observed between the pathways for the genes enriched within the differential regions as detected by Hi-C and the differentially expressed genes as detected by RNA-seq suggests that 3D DNA structure is indeed associated with gene expression.

Next, we checked if the DE genes from the RNA-seq analysis were enriched in the differential regions detected by HiCcompare using the multiHiCcompare permutation test functionality for checking enrichment of genomic features in differentially interacting regions. The DE genes were not significantly enriched in the differential regions compared to randomly selected regions of the genome in the 1MB, 100KB, and 20KB resolution results ($P = 0.096$, $P = 0.59$, $P = 0.29$, respectively). A reason for this could be that the RNA-seq data was not taken from the same samples used to generate the Hi-C data. Additionally, the Hi-C data was taken from the dorsolateral prefrontal cortex while the RNA-seq data was from just the frontal cortex.

Table 3.8 KEGG pathway enrichment results using differentially expressed genes from the RNA-seq analysis.

Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
Calcium signaling pathway	30/339	188/7840	0.0000000	0.0000001	0.0000001
Adrenergic signaling in cardiomyocytes	22/339	145/7840	0.0000002	0.0000297	0.0000223
Dopaminergic synapse	20/339	131/7840	0.0000007	0.0000643	0.0000482
Salivary secretion	15/339	90/7840	0.0000060	0.0003681	0.0002759
Glutamatergic synapse	17/339	114/7840	0.0000070	0.0003681	0.0002759
cAMP signaling pathway	24/339	212/7840	0.0000135	0.0005813	0.0004357
Circadian entrainment	15/339	97/7840	0.0000155	0.0005813	0.0004357
Aldosterone synthesis and secretion	15/339	98/7840	0.0000176	0.0005813	0.0004357
Type II diabetes mellitus	10/339	46/7840	0.0000203	0.0005949	0.0004459
GABAergic synapse	14/339	89/7840	0.0000245	0.0006457	0.0004840
Oxytocin signaling pathway	19/339	153/7840	0.0000301	0.0006963	0.0005219
Morphine addiction	14/339	91/7840	0.0000316	0.0006963	0.0005219
Taste transduction	13/339	83/7840	0.0000500	0.0010149	0.0007608
Neuroactive ligand-receptor interaction	31/339	338/7840	0.0000545	0.0010272	0.0007700
Axon guidance	20/339	181/7840	0.0001005	0.0017680	0.0013253
Arginine biosynthesis	6/339	21/7840	0.0001953	0.0032228	0.0024158
Retrograde endocannabinoid signaling	17/339	148/7840	0.0002080	0.0032303	0.0024214
cGMP-PKG signaling pathway	18/339	166/7840	0.0002820	0.0039324	0.0029477
Amyotrophic lateral sclerosis (ALS)	9/339	51/7840	0.0002891	0.0039324	0.0029477
Insulin secretion	12/339	86/7840	0.0002997	0.0039324	0.0029477

3.3.6 Amygdala and PFC share similar TAD structure

SpectralTAD detected a total of 10,970 TADs in the amygdala of which 2,927 were level 1 TADs, 4,098 were level 2, and 3,945 were level 3. For the PFC data, SpectralTAD detected a total of 10,833 TADs of which 2,912 were level 1, 4,044 were level 2, and 3,877 were level 3. Thus, the total number of TADs and TAD hierarchy lined up fairly well between the two regions of the brain. The Jaccard overlap between the TAD boundaries was found to be 0.549. Another study found that the average Jaccard overlap between different tissues was 0.416 (Natalie Sauerwald 2018), suggesting that the TAD structure of the amygdala and PFC are more similar than would be expected for most tissues.

To test if the differentially interacting regions detected by HiCcompare were enriched in TAD boundaries, we performed permutation tests to compare overlap of the differential regions to randomly selected regions. The regions detected in the 1MB resolution comparison were significantly enriched in the amygdala TAD boundaries (P = 0.019); however, they were not significantly enriched in the PFC TAD boundaries (P = 0.073). The 100KB differential regions were again found to be significantly enriched in the amygdala TAD boundaries (P = 0.003) but not in the PFC boundaries (P = 0.057). At 20KB resolution, the differential regions were not significantly enriched for either the amygdala or PFC TAD boundaries (P = 0.50 & P = 0.24).

3.4 Discussion

Using comparative analysis techniques, we detected several significantly different interactions in the 3D structure of the DNA between the amygdala and PFC. As expected, we found more differences between the amygdala and PFC samples than were found between the PFC replicates. The enrichment analysis of the differentially interacting regions revealed some biologically interesting results. We found several signaling pathways that were enriched within the DNA regions detected as differentially interacting. Activation of certain signaling pathways has been found to determine the fate of stem-cells (Blank 2008). As our data came from

fetal brains, the cellular differentiation process may still have been ongoing, which could be responsible for the enrichment of these signaling pathways.

The enrichment of the axon guidance pathway was also found in the study of Hi-C data from the human brain by Won et al. (Won et al. 2016). This suggests that the differential regions we found enriched for the axon guidance pathway is a true structural difference of the DNA influencing differentiation between the amygdala and PFC. Won et al. additionally found enrichment of cell adhesion and cytoskeleton protein binding as we did in our enrichment analysis. The RNA-seq differential gene expression analysis also produced DE genes enriched in several of the same pathways found in the Hi-C analysis. The results of our comparative analysis are consistent with those of this previous study, suggesting that `HiCcompare` can detect biologically relevant differences in the 3D structure of the DNA.

In our TAD analysis, we found that the TAD hierarchy and structure were consistent between the amygdala and PFC. The Jaccard overlap between TADs of the regions was higher than in other studies (Natalie Sauerwald 2018). Additionally, we found that the differential regions were significantly enriched in the amygdala TAD boundaries at 1MB and 100KB resolution, but not in the PFC TADs or either at 20KB resolution. The reason for no enrichment in 20KB resolution data could be due to the lower quality and sparsity of the data at this high of a resolution. It is also possible that many of the differences detected had to do with TAD boundaries changing in just the amygdala, thus explaining why significant enrichment was only found for these TAD boundaries and not in the PFC boundaries.

One main limitation of this study was the small sample size and the resolution of the data. As we had only one sample for the amygdala and two technical replicates of different sequencing quality for the PFC, our analyses were limited. The quality of the sequencing and low sample size limited the results of this study and prevented any broad conclusions. Additionally, as the sequencing depth was not sufficient, we could not make much use of the data at high-resolution. Further study of the 3D structure of brain DNA data should require more samples per tissue type at higher resolutions. Further study will help to characterize the regulatory nature of the DNA interactions in the brain differentiation process.

Chapter 4: Aim 3 - Joint normalization and difference detection for replicate Hi-C datasets

4.1 Introduction

The advent of Chromatin Conformation Capture (3C) technology allowed for the first insights into the three-dimensional (3D) interactome of the genome (Dekker et al. 2002b). Following 3C, 4C and 5C, a high-throughput technology, Hi-C, was introduced as a means for the capture of all vs. all interactions across the entire genome (Lieberman-Aiden et al. 2009b). The structure and interactions of the DNA in 3D space inside the nucleus has been shown to shape the gene expression of cells and define cellular identity (Dowen et al. 2014b; Ji et al. 2016b; Phillips-Cremins and Corces 2013b; Rao et al. 2014b; Vietri Rudan et al. 2015; Jin et al. 2013b) and in the regulation of tumor suppressors and oncogenes (Valton and Dekker 2016; Rickman et al. 2012b; Taberlay et al. 2016b; Hnisz et al. 2016b; Lupiáñez et al. 2016b). The dynamic nature of the 3D structure of the genome prompted significant attention to the comparative analysis of multiple Hi-C datasets (Dixon et al. 2015; Bonev et al. 2017).

Soon after Hi-C data became available, it became clear that the data contained biases which affected the construction and analysis of chromatin contact maps (Yaffe and Tanay 2011b). These biases fall into two categories: DNA sequence-driven and sequencing technology-driven. The *sequence-driven biases* that can be explicitly modeled include GC content, chromatin accessibility, nucleosome occupancy, repetitive elements and other properties of the DNA sequence (Yaffe and Tanay 2011b; O'Sullivan et al.b), and are consistent across datasets. The much less understood and hard-to-model *technology-driven biases* include cross-linking preferences, the choice of restriction enzymes (e.g., HindIII, MboI, DpnII), biotin labeling, chromatin fragmentation, and sequencing depth, among others. These biases affect Hi-C datasets unpredictably, justifying the need for joint normalization of multiple datasets. Early studies tended to focus on normalizing individual Hi-C datasets (Lieberman-Aiden et al. 2009b; Imakaev et al. 2012b; Yaffe and Tanay 2011b; Knight and Ruiz 2012b). These individual methods improve the reproducibility of replicated datasets (Imakaev et al.

2012b; Yaffe and Tanay 2011b; Hu et al. 2012b). However, these methods leave the problem of different biases between multiple Hi-C datasets unaddressed.

Early methods for normalizing and comparing Hi-C datasets were developed to normalize individual datasets and overlap them. The most notable example is the `HiCCUPS` algorithm (Rao et al. 2014b), which detects chromatin interaction “hotspots” in individually normalized Hi-C datasets. Hotspots, i.e., chromatin interactions enriched relative to the local background, are then compared between datasets by simply overlapping them. This approach does not distinguish hotspots detected due to local biases and does not quantify the significance of the differences. Several papers utilized individually normalized Hi-C datasets and overlap-based methods to reveal important insights into the dynamics of the 3D structure of the genome (Dixon et al. 2015; Bonev et al. 2017). However, the overlap-based methods are severely limited in detecting statistically significant chromatin interaction changes.

To the best of our knowledge, only four methods approach a statistically grounded comparison of Hi-C datasets. The `diffHic` method is an extension of a negative binomial distribution-based analysis operating on count data (Lun and Smyth 2015b). As such, it leaves a user with challenges of sequencing data storage, the computational burden of processing, normalization, summarization, and other bioinformatics heavy lifting of Hi-C data. The `HOMER` method uses a binomial model to compare individually normalized Hi-C datasets (Heinz et al. 2010). The `ChromoR` method (Shavit and Lio’ 2014b) uses a Poisson model to compare Hi-C datasets. The latest method, `FIND`, exploits a spatial Poisson process to consider spatial dependency between chromatin regions when detecting differentially interacting loci (Djekidel et al. 2018b). However, in our tests, these methods failed to detect consistent differential chromatin interactions. Furthermore, all but `diffHic` methods use individually normalized Hi-C datasets, leaving the technology-driven biases unaccounted for. Thus, the problem of normalization and statistical comparison of multiple Hi-C datasets remains unsolved.

Our method, `HiCcompare` (Stansfield et al. 2018), was one of the pioneering normalization methods to consider between dataset biases; however, it is limited to only joint normalization and comparison of two datasets. As sequencing costs continue to decrease and availability of Hi-C sequencing data increases, this method will fall short for Hi-C experiments involving comparison of multiple datasets.

We present a method, `multiHiCcompare`, for joint normalization and comparison of multiple Hi-C datasets. Our method is based on a distance-centric view of Hi-C data, accounting for the fact that chromatin interaction frequencies (IFs) decay with the increasing distance between interacting regions. Our method utilizes cyclic loess regression-based normalization to jointly normalize Hi-C datasets between replicates and conditions. We then present a differential chromatin interaction analysis framework based on a general linear model (GLM)-based approach (McCarthy et al. 2012; Robinson et al. 2010). Our framework operates on interaction counts subset in a distance-centric manner to produce RNA-seq-count-like matrices that can be directly analyzed using the GLM approach. As an output, genomic coordinates of differentially interacting regions are reported in text format, and the results are compatible with `Juicer` (Durand et al. 2016b) for easy visualization. This method, implemented in the `multiHiCcompare` R package, represents a streamlined and well-documented pipeline for the joint normalization and comparative analysis of multiple Hi-C experiments.

4.2 Methods

4.2.1 Hi-C data format

`multiHiCcompare` works on processed Hi-C data in the form of sparse upper triangular matrices, in plain text format. A typical sparse Hi-C matrix is stored in a separate file for each chromosome and contains three columns - the start location for the first interacting regions, the start location for the second interacting region, and the interaction frequency for that interaction. When importing data to use in `multiHiCcompare`, an additional column needs to be added indicating the chromosome number, as the first column. The original `HiCcompare` package provides functions for converting between full and sparse matrices (Stansfield et al. 2018).

4.2.2 Filtering

Pairs of chromatin regions showing zero interaction frequency (IF) across all samples are not considered in all analyses. Additional filtering options include filtering out interacting pairs of regions with the average IF below a pre-defined threshold and/or the proportion of zero IF values larger than a pre-defined threshold across multiple datasets. Filtering helps to increase the computational speed when normalizing and comparing the data. Additionally, it removes interactions with low variability and high numbers of zero IFs that may create problems when estimating the parameters of the negative binomial distribution in the comparative analysis step (Lun and Smyth 2017). Furthermore, filtering helps to increase power by reducing the effect of the multiple testing correction. By default, interaction pairs with an average IF less than 5 and the proportion of zero IFs larger than 80% are filtered out.

4.2.3 Cyclic loess normalization

We previously developed a loess regression-based method for normalizing two Hi-C datasets (Stansfield et al. 2018). Briefly, the method is based on representing the data on an MD plot. The MD plot is similar to the MA plot (Bland-Altman plot) (Dudoit et al. 2002b) which is commonly used for the visualization of gene expression differences. M is defined as the log difference between the two data sets $M = log_2(IF_2/IF_1)$, where IF_1 and IF_2 are interaction frequencies of the first and the second Hi-C datasets, respectively. D is defined as the distance between two interacting regions, expressed in unit-length of the resolution of the Hi-C data. A loess regression curve is fit through the MD plot and used to remove global biases by centering the M differences around $M = 0$ baseline (Stansfield et al. 2018). In our previous work, we show that joint loess normalization on the MD plot is superior to other common Hi-C normalization methods (ICE, KR, MA) for the purpose of comparison between experimental conditions (Stansfield et al. 2018). We also performed an additional comparison of cyclic loess with `HiCNorm` (Hu et al. 2012b) (Appendix 5 Figure S1).

Here, we adapt our method to normalizing multiple Hi-C datasets, a procedure termed “cyclic loess” (Ballman et al. 2004). The cyclic loess algorithm proceeds through the following steps:

1. Choose two out of the N total samples then generate an MD plot.
2. Fit a loess curve $f(d)$ to the MD plot.
3. Subtract $f(d)/2$ from the first dataset and add $f(d)/2$ to the second.
4. Repeat until all unique pairs have been compared.
5. Repeat until convergence.

Cyclic loess typically requires two to three iterations for convergence, which is reached when the fitted values stop changing (Ballman et al. 2004). The order in which pairs are selected may negligibly affect the speed of convergence; however, there should be no effect on normalization results. (Ballman et al. 2004).

Cyclic loess is computationally intensive; however, it naturally lends itself to parallelization. If a parallelization option is specified, normalization for each chromosome is distributed to a different processor. `multiHiCcompare` makes use of the BiocParallel parallelization methods.

Additionally, we implemented a modified version of the fast linear loess (“fastlo”) method (Ballman et al. 2004) that is adapted to Hi-C data on a per-distance basis. To perform “fastlo” on Hi-C data, we first split the data into p pooled matrices. “Progressive pooling” is used to split up the Hi-C matrix by unit distance such that distance 0 is its own pool, distances 1 and 2 are pooled, distance 3, 4, 5 are pooled, and so on until all unit distances belong to one of p pools. p is calculated as follows $p = ceiling\left\{\frac{\sqrt{8n_d+1}-1}{2}\right\}$ where n_d is the number of unit distances. The solution for the number of pools follows from the quadratic formula solution for triangular numbers. Progressive pooling is required for the fastlo and difference detection steps because each off-diagonal trace of the matrix gets progressively smaller than the last. Thus, progressive pooling allows for normalization and analysis to be performed in a distance-centric manner while maintaining a similar number of contacts in each pool. These pooled contacts are assembled into matrices of interaction frequencies. Each matrix will have an IF_{gj} value with g interacting pairs for each of the j samples. These p matrices can then be input into the “fastlo” algorithm using the following steps:

1. Create the vector $\hat{IF}_{pgj}$, the row means of the p^{th} matrix. This is the equivalent of creating an average IF at distance pool p.
2. Plot $\hat{IF}_p$ versus $(IF_{pg} - \hat{IF}_p)$ for each sample j. This is equivalent to an MA plot at a genomic distance pool p.
3. Fit a loess curve $f(x)$ to the plot.
4. Subtract $f(x)$ from sample j.
5. Repeat for all remaining replicates.
6. Repeat until the algorithm converges.

The above steps are performed on the log2-transformed IFs. If a parallelization option is specified, the "fastlo" algorithm is parallelized by splitting up the p matrices and sending them to multiple processors. Similarly to cyclic loess, fastlo typically converges within two iterations, which is defined as the point when the row means no longer change (Ballman et al. 2004). Additionally, fastlo has been shown to provide similar normalized values as quantile normalization while being almost as fast computationally (Ballman et al. 2004). Both the cyclic loess and fastlo methods are included in the `multiHiCcompare` package.

After joint normalization, any negative IFs are automatically set to values of 0. All IFs that started with a zero value are reverted to zero after normalization is complete. This is because we are unable to determine if zeros in Hi-C matrices represent a missing value or an actual absence of contact between the pair of regions.

4.2.4 Detection of chromatin interaction differences

After normalization of the data, we can then proceed to the differential analysis. The primary goal of the differential analysis is to detect the maximal number of true differences while minimizing false positives. Approaches that utilize information across replicate high-throughput data (microarrays, RNA-seq, ChIP-seq) have been shown to improve the power of differential analysis (Smyth 2004; Sartor et al. 2006; Phipson et al. 2016; Yu et al. 2011). Adopting the distance-centric view of Hi-C data (the off-diagonal vectors in chromatin interaction matrices, Figure 1), a comparison with other sequencing technologies can be drawn. Similar to RNA-seq read counts, Hi-C IFs may have differing amounts of biological variation across replicates.

As Hi-C reads forming pairwise interaction frequencies are count based, the IFs can be modeled using a Negative Binomial distribution (Robinson and Smyth 2007) (Appendix 5 Figure S2). The distributions of distance-centric vectors of interaction counts can be approximated by the NB distribution, and this approximation holds at different resolutions of Hi-C data and different distances between interacting regions. Thus, the general linear model (GLM) framework of differential gene expression analysis developed for RNA-seq (McCarthy et al. 2012; Auer and Doerge 2010; Anders and Huber 2010; Baggerly et al. 2003, 2004; Hansen et al. 2011; Lu et al. 2005; Robinson and Smyth 2007, 2008; Robinson et al. 2010) can be adapted for differential analysis of interaction frequencies. We adapted this framework to process interaction frequencies represented as p "progressively pooled" distance-centric matrices with g rows (indices for interacting pairs of regions) and i columns (indices for replicates, Figure 1). The "progressive pooling" strategy is aimed to increase the robustness of statistical estimates across the whole range of distances between interacting regions. Its adaptation for the GLM framework is described in the next section.

4.2.5 Adaptation of differential detection statistics for Hi-C data

To detect chromatin interaction differences between multiple Hi-C datasets, we define $j = 1,2$ as an experimental condition for which replicated Hi-C datasets were produced. The goal is to identify differential chromatin interactions between condition $j = 1$ vs. condition $j = 2$. The IF value for a specific interacting pair of regions, g, at distance pool, d, from sample i will be denoted $y_{dgi} \sim NB(M_{di}p_{dgj}, \phi_{dg})$, where M_{di} is the total number of reads in sample i at distance pool d, p_{dgj} is the proportion of interaction counts g in sample i from experimental condition j, and ϕ_{dg} is the dispersion. The NB distribution can be parameterized with the mean $\mu_{dgi} = M_{di}p_{dgj}$ and variance $var(y_{dgi}) = \mu_{dgi} + \phi_{dg}\mu_{dgi}^2$ (Robinson et al. 2010). Dividing by μ_{dgi}^2 obtains the total coefficient of variation $CV(y_{dgi})^2 = 1/\mu_{dgi} + \phi_{dg} = Technical\ CV^2 + Biological\ CV^2$ It follows that the biological CV is $\sqrt{\phi_{dg}}$. For difference detection, the parameters of interest are p_{dgj} and ϕ_{dg}.

The easiest way to obtain the NB variance parameter is to estimate it globally across all distance-specific interaction frequencies, i.e., set $\phi_{dg} = \phi_d$ (Robinson and Smyth 2008; Anders and Huber 2010; Hansen et al. 2011). The common dispersion can be estimated through the maximizing the likelihood function $L(\phi_d) = \frac{1}{G}\sum_{g=1}^{G} L_g\,(\phi_d)$, where G is the number of interaction counts at a given distance pool d. Intuitively, it allows for prioritizing interaction frequencies that behave consistently across replicates, hence representing the more reliable measures. This approach works well when the number of replicates is small (Robinson and Smyth 2007; Kal et al. 1999), as will be the case in most Hi-C experiments. An extension of this simplified approach is to consider the dependence between the variance function and the mean of IFs, so that all IFs with the same expected count have the same variance. Frequency-specific dispersions can then be estimated using an empirical Bayes approach using the quasi-likelihood (Baggerly et al. 2004; Tjur 1998), weighted likelihood (Robinson and Smyth 2007) or Cox–Reid adjusted profile likelihood (McCarthy et al. 2012), to be then used for moderation of outlier variances.

To test for differences in chromatin interactions between two cellular conditions a likelihood ratio test can be used. We are primarily interested in testing the null hypothesis $\mu_{dg,j=1} = \mu_{dg,j=2}$. The expense and sequencing depth required for generating Hi-C data means that the number of replicates for each condition will typically be very small. Thus, we estimate the common dispersion at distance pool d, ϕ_d, as described (Robinson and Smyth 2008; Anders and Huber 2010; Hansen et al. 2011). Using the estimated common dispersion for a distance pool d, we can then perform an exact test similar to Fisher's exact test (Robinson and Smyth 2008). The cyclic loess normalization will allow the IFs for interacting pairs to be treated as identically distributed negative binomial random variables. For each pairwise interaction g, at a distance pool d, we can calculate the total sum of experimental groups $j = 1{,}2$ and compare it to the group-wise totals to calculate a 2-sided p-value.

To analyze data from more complex experimental designs (three or more groups, covariates), generalized linear models (GLMs) have been extensively used for differential analysis of non-normally distributed count data (Tjur 1998; Baggerly et al. 2004; Lu et al. 2005; Auer and Doerge 2010; Anders and Huber 2010; McCarthy et al. 2012; Robinson et al. 2010). Assuming that an estimate is available for ϕ_{dg}, so the variance can be calculated for any μ_{dgj}, the vector of covariates (e.g., experimental condition assignment) x_i can be linked with μ_{dgj} through a log-linear model $log(\mu_{dgj}) = x_i^T \beta_{dg} + log(M_{dj})$, where β_{dg} is a vector of regression coefficients by which the covariate effects are mediated for interaction frequency g, and $log(M_{dj})$ is a log-transformed total number of reads in sample i at distance d that accounts for sequencing depth variability. β_{dg} can be estimated as $X^T z_{dg}$, where X is the design matrix with columns x_i and $z_{dgi} = (y_{dgi} - \mu_{dgi})/(1 + \phi_{dg}\mu_{dgi})$. Importantly, for Hi-C replicates generated using different enzymes the design matrix X may contain a covariate column for an enzyme, naturally accounting for the variability introduced through differences in DNA digestion. The GLM framework was implemented into `multiHiCcompare` and can be used to detect differential interactions from replicated Hi-C data.

4.2.6 Benchmarking `multiHiCcompare`

To accurately benchmark a method, data with ground truth differences are required (Dozmorov et al. 2010b). As there is no gold standard for differential interactions in Hi-C data, we used technical replicates from HCT-116 colorectal cancer cell line at 100KB resolution for chromosome 22 (Rao et al. 2017) to generate a set of 4x4 Hi-C matrices with ground truth differences. To create this dataset, we used four technical replicates ("Normal; Biological Sample 2") (Appendix 5 Table S1) and created an additional four Hi-C datasets by adding random noise to each of them. Noise was estimated by fitting the distributions of the differences between the replicate dataset's IFs. The differences were found to follow a roughly normal distribution with means near 0 and standard deviations between 8 and 11. Thus, to add noise to our "simulated" replicates, we sampled from a normal distribution with mean 0 and standard deviation of 10. The noise matrix was then added to the real Hi-C data to produce the simulated replicates. This created a total of eight semi-simulated replicate datasets, suitable for the 4x4 group comparison.

A pre-specified number of ground truth differences were added in randomly to the chromatin interaction matrices. The randomly selected interacting pairs had their IFs set to the mean of all samples, and then Gaussian noise sampled from a $N(0, \sigma)$ distribution was added to the IFs. σ is defined by fitting a linear

regression between average IF and standard deviation of IFs. Finally, the IFs for one of the experimental conditions were multiplied by a pre-specified fold change. This method produces an average fold change difference between the conditions while still preserving some variation in the IFs from different samples.

To illustrate the benefits of replicated Hi-C data, two parameters of comparative analysis were tested - 1) the number of replicates, and 2) the fold change. Additionally, we investigated the effect of the resolution of Hi-C data (finer resolution is expected to have a lower dynamic range and the higher proportion of zero IFs). We also compared the performance of `multiHiCcompare` with the original `HiCcompare` method (Stansfield et al. 2018). Using the ground truth differences as a reference, we performed an ROC analysis as well as assessed other standard performance classifiers.

4.2.7 Comparison with `FIND`

To compare the performance of `multiHiCcompare` with that of `FIND`, we used `FIND`'s `generateSimulation()` simulation function to generate datasets with varying numbers of replicates for two conditions. A 300x300 Hi-C contact map with a 0.002 probability of differential interactions at fold changes of 3, 5, and 10 were generated for each replicate. Differential chromatin interactions detected by `FIND` and `multiHiCcompare` were compared against the ground truth, and the ROC analysis and standard performance classifiers were assessed.

4.2.8 Comparison with `diffHic`

To compare the performance of `multiHiCcompare` with that of `diffHic`, we performed an analysis of the Hi-C data used in the `diffHic` paper (Lun and Smyth 2015b). The data was from human prostate epithelial cells RWPE1 over-expressing the ERG protein or GFP protein, generated in replicates (Rickman et al. 2012b) (Appendix 5 Table S1). The sequencing data were processed into BAM files using `HiCUP` (Wingett et al. 2015) and then converted to contact maps using `juicer tools` (Durand et al. 2016b). The contact maps for the Hi-C libraries were then inputted into `multiHiCcompare`. The data were jointly normalized with cyclic loess, and the exact test was used to detect differences between the two conditions. The results of the `multiHiCcompare` analysis were then compared with those from the `diffHic` analysis.

4.2.9 Comparison of auxin-treated vs. untreated Hi-C datasets

As a case example of `multiHiCcompare`, we performed a differential analysis of Hi-C data from (Rao et al. 2017). The data from HCT-116 colon carcinoma cell line had 7 samples from 2 biological sources for the untreated condition and 7 samples from 2 biological sources for the auxin treatment group (Appendix 5 Table S1). We jointly normalized all replicates using `multiHiCcompare`. We then input the data into the GLM framework to detect differences between the two conditions while controlling for the biological source. The regions that were detected as differentially interacting were assessed for enrichment in HCT-116-specific transcription factor binding sites (Appendix 5 Table S3).

4.2.10 Analysis of CTCF depleted cells

As an additional example of the functionality of `multiHiCcompare`, we performed a differential analysis of Hi-C data from (Jessica Zuin 2014) at 40KB resolution. The experiment was performed on HEK293 cells which had CTCF siRNA knockdowns compared to control cells. Two control samples and two CTCF knockdown samples (Appendix 5 Table S1) were normalized and compared using `multiHiCcompare`. The regions that were detected as differentially interacting were assessed for enrichment in HEK293 specific transcription factor binding sites (Appendix 5 Table S5).

4.2.11 Software availability

`multiHiCcompare` is freely available as an R package on Bioconductor at https://bioconductor.org/packages/multiHiCcompare and Github at https://github.com/dozmorovlab/multHiCcompare. The package includes a vignette and test data along with documentation for all functions. `multiHiCcompare` is released under the MIT open source software license.

4.2.12 Data Access

For our benchmarking of `multiHiCcompare`, we used 14 samples from HCT-116 human colorectal carcinoma cell line (Rao et al. 2017). For the comparison with `diffHic`, we used data from RWPE1 prostate cancer epithelial cell lines over-expressing the ERG protein or GFP protein (Rickman et al. 2012b). For the enrichment analysis of differentially interacting regions in HCT-116 and HEK293 cells, we used ChIP-seq transcription factor binding sites from CistromeDB (Mei et al. 2017). All data sources are presented in Appendix 5 Tables S1, S3, and S5.

4.3 Results

4.3.1 multiHiCcompare method outline

`multiHiCcompare` is an R package for the joint normalization and detection of chromatin interaction differences in multiple Hi-C datasets. A basic `multiHiCcompare` analysis will start with pre-processed Hi-C data from two or more experimental conditions for which each condition has one or more samples (technical or biological replicates). The whole-genome Hi-C data should be provided as a single file in the form of plain text four column sparse upper triangular matrices. The data is then jointly normalized using either our cyclic loess or fastlo methods. Finally, the experimental conditions can be compared using either an exact test or a generalized linear model (GLM) framework, depending on the complexity of the experimental design. The flowchart in Figure 4.1 shows a typical `multiHiCcompare` workflow.

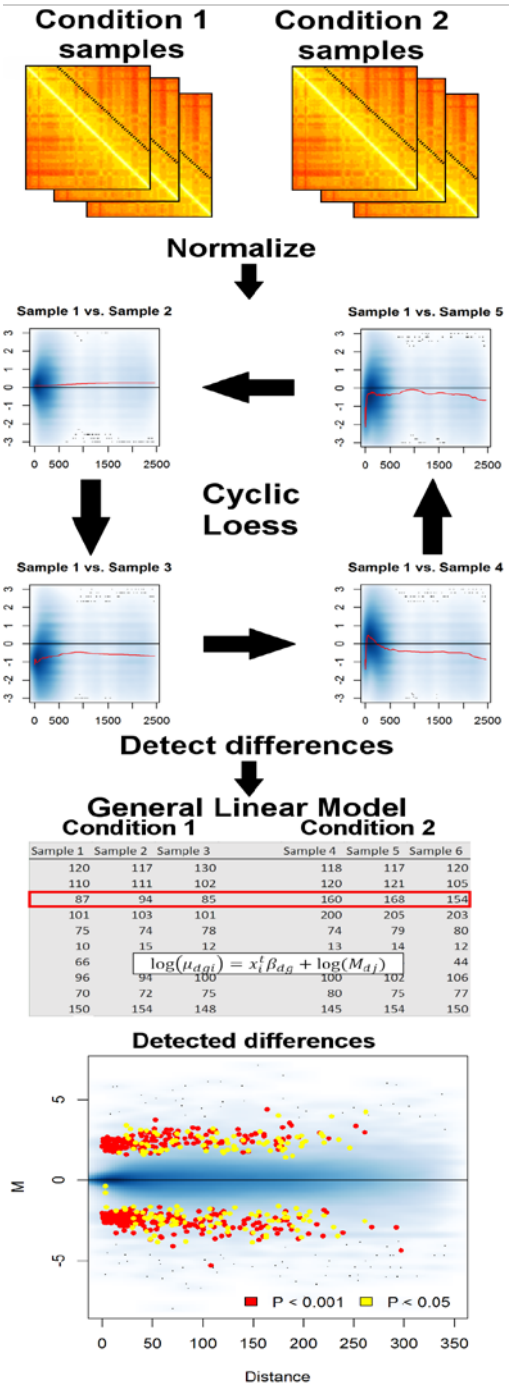

Figure 4.1. Flowchart for a multiHiCcompare analysis. Pre-processed Hi-C data is read in and then normalized using the cyclic loess (or fastlo) methods. Then, “progressive pooling” of the off-diagonal (distance-centric) IFs into a matrix format is performed for input into either an exact test or GLM. Finally, the results of the comparison are shown on a composite MD plot indicating where the differences occurred.

4.3.2 Replicates of Hi-C data improve the power of detection of differential chromatin interactions

The performance of `multiHiCcompare` was quantified by using varying numbers of replicates per condition with added true differences at varying fold changes (see Methods). `multiHiCcompare` was able to detect the majority of the introduced differences with relatively low numbers of false positives, and the power of detecting differential interactions increased dramatically as the number of replicates in each experimental condition and the fold change increased (Figure 4.2). These results emphasize the utility of the GLM for differential chromatin interaction analysis.

The performance of multiHiCcompare was also tested against the original HiCcompare, which is designed to compare two datasets. We found that both methods performed well in detecting the added differences; however, HiCcompare had a larger area under the ROC curve in cases with one replicate per experimental condition (Figure 4.2). This is likely due to the limitations in calculating the dispersion factor for the negative binomial model used in multiHiCcompare when no replicates are available. Therefore, for 1x1 dataset comparison, we recommend using the original HiCcompare method, while when multiple replicates are available multiHiCcompare is more powerful at detecting true differences.

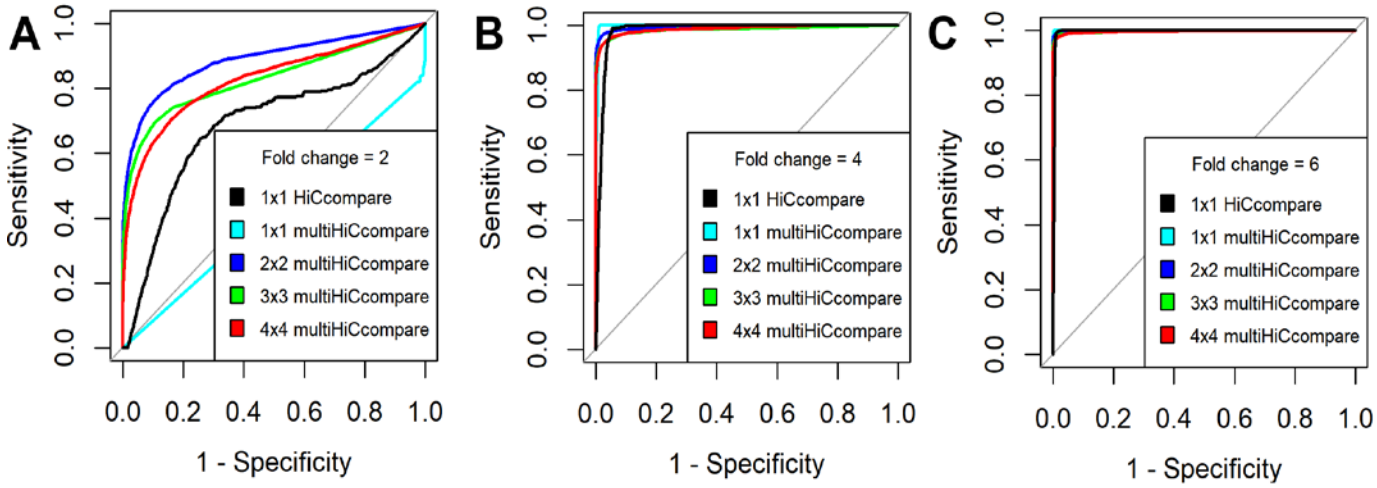

Figure 4.2. ROC analysis of the performance of multiHiCcompare and HiCcompare over various fold changes for introduced differences. The ROC curves demonstrate the increase in power in detecting differential chromosome interactions as the number of replicates per experimental condition increases from 1 to 4 compared with the performance of HiCcompare at 2, 4, 6-fold changes, panels A, B, C, respectively.

4.3.3 multiHiCcompare outperforms FIND

To compare the performance of multiHiCcompare with FIND, a recently published method for differential chromatin interaction detection (Djekidel et al. 2018b), we generated simulated Hi-C matrices with true differences at 2, 4, and 6-fold changes. To test the effect of the number of replicates for each of these fold changes, we performed 2x2, 3x3, and 4x4 analyses. We found that over the range of fold changes, multiHiCcompare detected more true positives with less false positives than FIND (Appendix 5 Table S2) and showed a larger area under the ROC curve performance (Appendix 5 Figure S3). The Matthews Correlation Coefficient (MCC) for multiHiCcompare was also higher than that of FIND for the majority of tested examples (Figure 4.3). At the higher fold changes tested, multiHiCcompare was able to detect nearly two times the amount of true differences compared to FIND (Appendix 5 Table S2). These results demonstrate that multiHiCcompare outperforms FIND in the detection of differential chromatin interactions at different fold changes and different numbers of replicates.

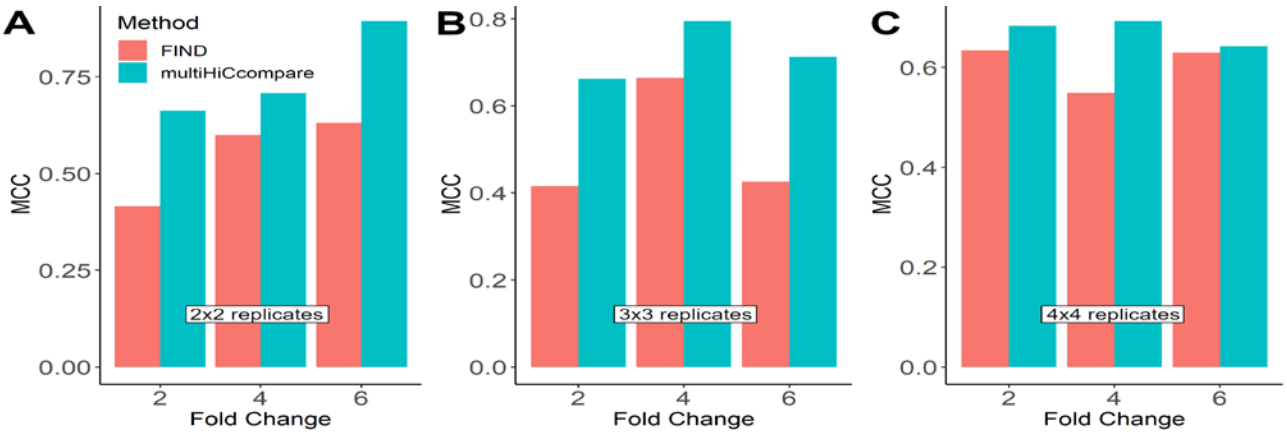

Figure 4.3. Comparison of Matthews Correlation Coefficient (MCC) between `multiHiCcompare` and `FIND` over various fold changes and 2x2, 3x3, and 4x4 numbers of replicates per condition, panels A, B, C, respectively. $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$.

4.3.4 `multiHiCcompare` identifies similar chromatin interaction differences detected by `diffHic`

We compared the performance of `multiHiCcompare` with the `diffHic` method (Lun and Smyth 2015b). We used the Hi-C data from human prostate epithelial cells (RWPE1 cells) overexpressing the ERG protein or a GFP control, analyzed in the `diffHic` paper (Rickman et al. 2012b). `multiHiCcompare` found 1,752 differences (FDR < 0.05) between the ERG and GFP conditions, more than was found in the original `HiCcompare` analysis (Stansfield et al. 2018), yet less than the 5,737 differences (FDR < 0.05) detected by `diffHic`. As shown in Figure 4.2, `multiHiCcompare` might have gained some additional power over `HiCcompare` by making use of the two Hi-C libraries (the `multiHiCcompare` analysis was a 2x2 analysis, compared to the 1x1 analysis of `HiCcompare`). However, both `HiCcompare` and `multiHiCcompare` seem to be more conservative than `diffHic`. The overlap between the `multiHiCcompare`-detected and `diffHic`-detected differences was significant (1,254 overlapping regions, Fisher's exact test p-value < $2.2x10^{-16}$). This overlap is expected as both methods utilize the same GLM framework, while `multiHiCcompare` applies it with respect to the distance between interacting regions. Additionally, `multiHiCcompare` was able to detect all but one differential interactions validated by Fluorescence In Situ Hybridization (FISH) (Table 4.1), further confirming the power of `multiHiCcompare` in detecting biologically relevant chromatin interaction differences.

Table 4.1. Differential interaction statistics from `multiHiCcompare` and `diffHic` for chromatin interaction differences experimentally validated by FISH. "Interaction" - the genes interacting identified by FISH, "logFC" - the log2 fold change of interaction frequency difference between conditions, "logCPM" - the between-conditions average log counts per million of the IFs, for `multiHiCcompare` and `diffHic` results, respectively.

	multiHiCcompare				diffHic			
Interaction	logFC	logCPM	p-value	FDR	logFC	logCPM	p-values	FDR
FYN - MOXD1	-2.113	10.093	<0.001	0.007	0.733	1.134	0.002	0.042
HEY2 - MOXD1	1.232	11.182	<0.001	<0.001	0.67	2.625	<0.001	0.002
SERPINB9 - MOXD1	-2.227	9.621	0.008	0.356	-1.27	-0.151	0.001	0.016
FYN - HEY2	-2.113	10.093	<0.001	0.007	-1.545	0.621	<0.001	<0.001

4.3.5 `multiHiCcompare` is robust to the resolution of Hi-C data

Typically, Hi-C data at higher resolution (smaller size of chromatin regions tested for interactions) have a lower dynamic range and a higher proportion of zero IFs (sparsity). To examine the effect of resolution on the performance of `multiHiCcompare`, we calculated the Matthew's Correlation Coefficient (MCC) at resolutions of 50KB, 10KB, and 5KB (Appendix 5 Figure S5). `multiHiCcompare` encountered some difficulties at detecting

the added in differences at 2-fold changes in the very high-resolution data. However, at fold changes of 4 or greater, `multiHiCcompare` performed well at all resolutions. Evaluation of other performance metrics confirmed this conclusion (Appendix 5 Table S4). These results indicate that sparsity of the Hi-C data might hinder the detection of small differences at high-resolution, but overall `multiHiCcompare` appears to perform well even in sparse conditions.

4.3.6 `multiHiCcompare` detects regions associated with loss of chromatin loops in auxin-treated cells

We compared data from HCT-116 cells treated with auxin to those not treated (Rao et al. 2017). The auxin treatment is thought to eliminate chromatin loops, thus changing many chromatin interactions. The untreated group contained seven samples from biological replicates 1 and 2. The auxin-treated group contained seven samples from biological replicates 1 and 2 treated with auxin for 6 hours (Appendix 5 Table S1). All samples were jointly normalized, and differentially interacting chromatin regions were detected. The biological replicate number was entered as a covariate, and the main effect of auxin treatment was evaluated. This analysis was aimed at identifying regions associated with loss of chromatin loops.

We found a total of 417,145 differentially interacting pairs between the normal cells and the auxin treatment (FDR < 0.05). The auxin treatment is known to destroy the RAD21 protein of the cohesin complex and thus degrade chromatin looping. Therefore, we hypothesized that the regions detected by `multiHiCcompare` as differentially interacting should be enriched with RAD21 binding sites, and their interaction frequency should be decreased in auxin-treated condition. To test the significant differentially interacting regions for the enrichment of transcription factor binding sites, we performed permutation tests where a random set of genomic regions of the same size as the significant regions were sampled and compared for enrichment against the significant regions. Analysis of the most significant differentially interacting regions (FDR < 10^{-15}) showed that they were significantly enriched for RAD21 binding sites (permutation p-value < 0.001, Appendix 5 Table S3). Additionally, the regions enriched for RAD21 mostly exhibited lower IF values compared to the normal cells. Notably, we detected SMC1A, another structural maintenance protein of the cohesin complex reported to be affected by the auxin treatment (Rao et al. 2017), to be enriched in these regions (permutation p-value = 0.04). Consistent with the original findings, SMC1A enriched regions also exhibited lower IF values compared to the normal cells (Appendix 5 Table S3). Further, consistent with the original findings, we found that HCT-116 cell-specific CTCF sites were not enriched in the detected regions. These results indicate that `multiHiCcompare` is capable of detecting biologically relevant differences in chromatin conformation between experimental conditions.

In addition to the expected decrease in RAD21 and SMC1A binding sites, and no change in CTCF binding, we tested whether regions differentially interacting in auxin-treated condition are enriched in other HCT-116-specific transcription factors (Appendix 5 Table S3). The rationale here was to detect other transcription factors that may be responsible for chromatin loop formation. Notably, we detected strong enrichment in TCF4 binding sites (Table 4.2), a transcription factor previously linked to SMC3, a known component of the cohesin complex (Ghiselli et al. 2003). Furthermore, we observed enrichment of the heterochromatin protein HP1γ (also known as CBX3) and other proteins responsible for chromatin structure (Table 4.2). Expectedly, chromatin interaction frequency was decreased in these regions, confirming that auxin treatment leads to loss of chromatin loops formed by the cohesin complex (Rao et al. 2017). These findings confirm that `multiHiCcompare` allows for deeper insights into the biology of differential chromatin interactions.

Table 4.2. Transcription factors (TFs) significantly (p-value <0.05) enriched in the differentially interacting regions in HCT-116 auxin-treated cells. The Stouffer-Liptak method of combining p-values (Stouffer 1949) was used to obtain a summary p-value for each TF, as many TFs were represented by multiple datasets. “Number of experiments” - the number of ChIP-seq tracks supporting the enrichment, “Mean logFC” - the between-conditions average log fold change of regions overlapping with a transcription factor, “Stouffer-Liptak p-value” - enrichment p-value summarized using Stouffer-Liptak method (sorted by).

Transcription Factor	Number of experiments	Mean logFC	Stouffer-Liptak p-value
TCF4	8	-1.07	1.38E-07
CBX3	7	-0.60	2.95E-04
EP300	1	-0.89	9.99E-04
FOSL1	1	-0.89	9.99E-04
CEBPB	1	-0.87	9.99E-04
JUND	1	-0.87	9.99E-04
RAD21	1	-0.87	9.99E-04
KMT2B	1	-0.84	9.99E-04
SRF	1	-0.83	9.99E-04
TCF7L2	1	-0.83	9.99E-04
MAX	1	-0.82	9.99E-04
TEAD4	1	-0.82	9.99E-04
USF1	1	-0.79	2.00E-03
ATF3	1	-0.79	3.00E-03
ZBTB33	1	-0.58	4.00E-03
ZC3H8	1	-0.73	6.99E-03
YY1	1	-0.78	1.10E-02
ELF1	1	-0.76	1.10E-02
EGR1	1	-0.75	2.10E-02
SMC1A	1	-0.85	4.40E-02
SP1	5	-1.13	4.41E-02
AFF4	7	-0.19	4.87E-02
MECP2	2	-0.85	4.92E-02

We further hypothesized that the differentially expressed (DE) genes detected in (Rao et al. 2017) would be enriched within the regions detected by `multiHiCcompare` as differentially interacting. The list of DE genes was obtained from GEO (GSE106886) and matched with the corresponding regions by genomic coordinates. The DE genes (FDR < 0.05) were checked for enrichment within the most significant differentially interacting regions (FDR < 10^{-15}). We found that these genes were significantly enriched within the regions detected by `multiHiCcompare` (permutation p-value = $3.9 * 10^{-4}$). In summary, these results demonstrate that `multiHiCcompare` is a powerful tool to detect biologically relevant chromatin interaction differences.

4.3.7 `multiHiCcompare` detects regions associated with siRNA knockdown of CTCF

Similar to the analysis performed on the auxin-treated cells, we used `multiHiCcompare` to analyze an experiment of CTCF siRNA knockdown in HEK293 cells (Jessica Zuin 2014). CTCF is thought to play a role in shaping the 3D organization of the genome, especially in relation to topologically associated domains (TADs) (Phillips and Corces 2009; Vietri Rudan et al. 2015), and its knockout led to the reduction of intra-domain interactions with the concurrent increase in inter-domain interactions. Thus, it was expected that knockdown of

CTCF should lead to changes that can be detected by `multiHiCcompare`. We detected a total of 640 (FDR < 0.05) differences between the control and CTCF siRNA knockdown cells. 448 (70%) of the differentially interacting regions had positive fold changes (mean log fold-change 2.8), potentially reflecting the increased inter-domain interactions.

Knockdown of CTCF is expected to “free” its binding sites from the insulator effect of CTCF and allow the associated chromatin regions to interact. Indeed, the original study found that the promoters of genes differentially expressed after CTCF knockdown were enriched in CTCF binding sites (Jessica Zuin 2014). Analysis of the significant differentially interacting regions detected by `multiHiCcompare` showed that they were significantly enriched for CTCF binding sites (permutation p-value < 0.001, Table 4.3, Appendix 5 Table S5). Members of cohesin complex were also found to be enriched following CTCF knockdown (Jessica Zuin 2014); consequently, SMC3 member of cohesin complex was also found to be enriched in the differential regions (permutation p-value < 0.001, Table 3). These findings mirror the original results (Jessica Zuin 2014), further confirming that `multiHiCcompare` can detect known biological differences in Hi-C data.

We also detected strong enrichment of POLR2A binding sites in the differential regions identified by `multiHiCcompare`, not reported in the original study. Notably, upregulation of Polymerase genes, including POLR2A, following knockdown of TFII-I, an interacting partner of CTCF, has been reported (Marques M 2015). Their results suggest that the increase in inter-domain interactions followed by CTCF depletion is likely accompanied by an increase in transcription driven by RNA Polymerase II. In summary, these results suggest that `multiHiCcompare` can detect known and new findings in the comparative analysis of Hi-C data.

Table 4.3. Transcription factors significantly (p-value <0.05) enriched in the differentially interacting regions in HEK293 CTCF knockdown cells.

Transcription Factor	Number of experiments	Mean logFC	Stouffer-Liptak p-value
POLR2A	32	2.5403	6.87E-68
FOXM1	12	2.6438	1.38E-26
XRN2	9	2.6961	1.77E-20
TRIM28	8	3.1365	3.34E-18
MYC	9	3.674	1.07E-12
CDK9	3	2.2834	4.33E-08
CTCF	3	2.7039	4.33E-08
EP300	3	2.913	4.33E-08
TET3	3	2.7027	4.33E-08
BRD4	5	3.2043	6.64E-08
ELK4	3	1.0881	8.45E-08
WDR5	3	3.6028	8.45E-08
BAHD1	3	3.1755	6.19E-06
CREBBP	2	2.7888	6.19E-06
JMJD6	2	2.4619	6.19E-06
NCOR1	2	3.2093	6.19E-06
BRD1	4	3.2242	1.07E-05
C17orf96	2	4.3094	1.16E-04
RYBP	2	0.5388	5.42E-04
FANCD2	2	3.3978	7.63E-04
EMX1	2	3.5324	7.82E-04
BRD2	1	2.7475	9.99E-04
BRD3	1	1.4877	9.99E-04

CBX4	1	1.1606	9.99E-04
DCP1A	1	2.7263	9.99E-04
DDX21	1	2.7353	9.99E-04
DPPA2	1	3.1886	9.99E-04
HCFC1	1	2.5906	9.99E-04
PHF8	1	2.3149	9.99E-04
SMC3	1	2.7747	9.99E-04
TBL1X	1	2.2675	9.99E-04
TET2	1	3.0843	9.99E-04
TTF2	1	2.6618	9.99E-04
ZNF143	1	3.3148	9.99E-04
ZNF263	1	2.9817	9.99E-04
RNF2	2	-0.374	1.15E-03
BMI1	2	3.2265	4.97E-03
C10orf12	2	4.3306	2.50E-02
PCGF6	1	5.5337	3.10E-02
AFF4	1	0.9309	3.50E-02

4.3.8 Runtime evaluation

In our testing, both cyclic loess and fastlo normalization methods perform reasonably equally in regards to difference detection (Appendix 5 Figure S4); however, fastlo offers quicker computational speeds (Appendix 5 Figure S6A). We provide cyclic loess method as a conceptually straightforward and illustrative algorithm of the joint normalization of multiple datasets and recommend fastlo as the default joint normalization method.

When compared to `FIND`, `multiHiCcompare` showed a much faster runtime. We found that `FIND` was extremely slow on any Hi-C matrices that were relatively complete (low proportion of zeros). For example, at resolutions of 20KB - 50KB `FIND` runtimes were more than 72 hours, while `multiHiCcompare` can perform a comparable analysis in under 10 minutes (Appendix 5 Figure S6A). Thus, `multiHiCcompare` represents a fast and scalable method for joint normalization and detection of chromatin interaction differences.

The memory footprint expectedly increased with the increased resolution of the data and the number of replicates (Appendix 5 Figure S6B). However, the memory footprint depends on the sparsity of the data; hence, the high-resolution data may take less memory due to the increased sparsity. In summary, the whole-genome Hi-C data analysis can be performed on a desktop computer.

4.4 Discussion

As Hi-C datasets begin to be generated in multiple replicates, methods for the joint analysis of them are becoming crucial. Our methods address this need by providing a software implementation for the joint normalization of multiple datasets and the detection of differential chromatin interactions. As with any sequencing technologies, Hi-C data are unpredictably affected by technological biases, hindering the detection of chromatin interaction differences. While methods for normalization of individual Hi-C datasets have been developed (Lieberman-Aiden et al. 2009b; Imakaev et al. 2012b; Yaffe and Tanay 2011b; Knight and Ruiz 2012b), methods for joint normalization and comparative analysis of Hi-C data remain immature. We present the first method for jointly normalizing multiple Hi-C datasets by extending our `HiCcompare` loess regression-based method (Stansfield et al. 2018) and adapting the GLM-based difference detection method (McCarthy et al. 2012; Robinson et al. 2010) for the comparative analysis of multiple Hi-C datasets. `multiHiCcompare` can detect *a priori* known changes in replicate data with a low rate of false positives, and its power only increases with the increasing number of Hi-C replicates. We demonstrate that `multiHiCcompare` can detect biologically relevant regions associated with loss of chromatin loops in auxin-treated cells (Rao et al. 2017) and CTCF

knockdown cells (Jessica Zuin 2014) . We believe that if replicates of Hi-C data are available, they should be used in multiHiCcompare to gain the most power in detecting chromatin interaction differences.

The diffHic method (Lun and Smyth 2015b) pioneered the use of the negative binomial distribution and the GLM framework, originally implemented in the edgeR package (Robinson et al. 2010), for the comparative analysis of two Hi-C datasets. Other tools, such as HiBrowse (Paulsen et al. 2014), diffloop (Lareau and Aryee 2018), also utilized this framework. We further confirm the suitability of the negative binomial distribution for Hi-C data modeling (Appendix 5 Figure S2) and extend the edgeR functionality with the distance-centric view of Hi-C data. Our previous results (Stansfield et al. 2018) and the current implementation demonstrate that the distance-centric analysis of Hi-C data is a powerful approach to detect true chromatin interaction differences.

Interestingly, when comparing multiHiCcompare against FIND, multiHiCcompare performed much better than FIND even when using FIND's simulation function. This may be because FIND excels at detecting large fold changes (e.g., 10-fold or 20-fold changes) (Djekidel et al. 2018b), while multiHiCcompare performs well at fold changes as small as 2. Thus, besides being much faster than FIND (see "Runtime evaluation" results), multiHiCcompare is better suited for the detection of chromatin interaction differences across the whole range of fold changes.

In comparison with diffHic, multiHiCcompare showed similar performance in our analysis of the RWPE1 data. Although multiHiCcompare detected a smaller number of differences than diffHic, there was a significant overlap in the detected lists of regions. This is expected as both multiHiCcompare and diffHic use the GLM framework for difference detection but differ in the normalization approach and distance-based considerations implemented in multiHiCcompare. We feel that the distance-centric approach for joint normalization and difference detection, as implemented in multiHiCcompare, is better suited for the analysis of multiple Hi-C datasets.

In summary, the multiHiCcompare R package provides user-friendly methods for the joint normalization and comparative analysis of multiple Hi-C datasets. Our methods have been shown to perform similarly or better than other available methods. To date, multiHiCcompare is the only method for the joint normalization of multiple Hi-C datasets, which has been shown to outperform the commonly used methods for normalizing individual datasets (Stansfield et al. 2018). Finally, since multiHiCcompare is designed as a Bioconductor R package, it can be easily installed and used on all operating systems.

Chapter 5: Discussion

5.1 Conclusions

In this dissertation, we have outlined novel approaches for the normalization and comparison of Hi-C data. Additionally, our analysis of human brain data has provided some novel insights into the links between 3D chromatin interactions, gene expression, regulation, and tissue differentiation. In Chapter 2, we introduce a novel method for the normalization and comparison of two Hi-C datasets. HiCcompare represented one of the first methods for specific joint normalization and direct statistical comparison of Hi-C data between experimental conditions. HiCcompare is, however, limited in that it can only compare two Hi-C datasets at a time even if larger sample sizes are available. The R package, now published on Bioconductor, provides the scientific community a user-friendly method for the comparison of two Hi-C datasets.

In Chapter 3, we performed a comparative analysis on Hi-C data from two regions of the human brain. This analysis provides some novel insights into how differences in the 3D structure of the genome may influence gene expression and tissue differentiation in humans. We linked genes located within the highly differentially interacting regions with pathways relevant to brain function. Several of these pathways also overlapped with pathways enriched by the differential genes from the RNA-seq analysis. Additionally, the differential regions were found to be enriched in the amygdala TADs indicating that a change in TAD boundaries may be associated with tissue differentiation.

Chapter 4 presents the extension and generalization of our methods developed in Chapter 2 for working with Hi-C experiments that have larger sample sizes and other covariates. multiHiCcompare represents a full statistical framework for which to jointly normalize and compare complex Hi-C experiments. multiHiCcompare has also been published as a Bioconductor R package thus providing the community with an intuitive software package for analyzing complex Hi-C experiments. Additionally, we have written an R tutorial paper for Current Protocols in Bioinformatics (John C. Stansfield 2019) which provides detailed step-by-step documentation for how to use multiHiCcompare in an example analysis along with several downstream interpretation steps.

The methods developed in this dissertation provide a much-needed tool set for researchers working with Hi-C data. There are few available methods for the joint normalization and comparison of Hi-C data. Of the few currently available tools for comparing Hi-C data, many of them suffer from poor documentation and long run times. The tools presented here have been extensively documented and made with user-friendliness in mind. We hope that our methods developed here will become widely used in the study of chromatin conformation.

5.2 Future work

It is currently unknown how the properties of Hi-C data, e.g., resolution, distance, sparsity influence the power of a comparative analysis of the 3D structure of the genome. To design more effective experiments using Hi-C data, it will be essential to determine how these factors influence the power of a comparative analysis. A goal of future work should be to establish general guidelines for comparative Hi-C experiments such as the sample sizes and levels of sparsity required to achieve 80% power.

In this future work, the sparsity of many Hi-C datasets should be quantified at different resolutions. The rate of false positives will also need to be calculated in technical and biological replicates. This will give a baseline level of false positives to be expected when analyzing data. Analysis of technical and biological replicates can also further help understand how the variance of Hi-C data differs between different biological samples compared to technical replicates.

Finally, power will need to be estimated over varying experimental conditions. This can be achieved using technical and biological replicate data and introducing true differences at specified fold changes similar to the methods described in Chapters 2 and 4. Performing these comparisons over various sample sizes will allow us to estimate the required sample size to reach 80% power. We will also be able to estimate the power of an experiment given a predetermined sample size, sparsity, fold change of expected differences, etc.

Appendix

Appendix 1 Supplementary materials for Chapter 2: Aim 1. This PDF file contains supplemental methods (Section 1), a computation performance evaluation of `HiCcompare` (Section 3), additional validation of methods used in `HiCcompare`, and extended comparisons with `diffHic` and `FIND` (Section 6 & 7).

Appendix 2. Table of gene enrichment results for ESC vs neuron. This excel file contains a worksheet for the GO MF, GO BP, and KEGG pathway analysis results for the gene enrichment analysis between the ESC and neuron discussed in the results section.

Appendix 3. Table of gene enrichment results for ESC vs NPC. This excel file contains a worksheet for the GO MF, GO BP, and KEGG pathway analysis results for the gene enrichment analysis between the ESC and NPC discussed the in the results section.

Appendix 4. Table of gene enrichment results for NPC vs Neuron. This excel file contains a worksheet for the GO MF results for the gene enrichment analysis between the NPC and Neuron. The GO BP and KEGG pathway analysis did not return any significant results and thus are not included here.

Appendix 5. Supplemental figures and tables for results of `multiHiCcompare`.

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Auer PL, Doerge RW. 2010. Statistical design and analysis of rna sequencing data. *Genetics* **185**: 405–16.

Baggerly KA, Deng L, Morris JS, Aldaz CM. 2003. Differential expression in sage: Accounting for normal between-library variation. *Bioinformatics* **19**: 1477–83.

Baggerly KA, Deng L, Morris JS, Aldaz CM. 2004. Overdispersed logistic regression for sage: Modelling multiple groups and covariates. *BMC Bioinformatics* **5**: 144.

Ballman KV, Grill DE, Oberg AL, Therneau TM. 2004. Faster cyclic loess: Normalizing rna arrays via linear models. *Bioinformatics* **20**: 2778–86.

Battulin F N. 2015. Comparison of the three-dimensional organization of sperm and fibroblast genomes using the hi-c approach. *Genome Biology, 16(1), 77*.

Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. 2012. Hi-c: A comprehensive technique to capture the conformation of genomes. *Methods* **58**: 268–76.

Bernstein BE, Meissner A, Lander ES. 2007. The mammalian epigenome. *Cell* **128**: 669–81.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**: 6–21.

Birdsill AC, Walker DG, Lue L, Sue LI, Beach TG. 2011. Postmortem interval effect on rna and gene expression in human brain tissue. *Cell Tissue Bank* **12**: 311–8.

Blank K U. 2008. Signaling pathways governing stem-cell fate. *Blood, 111(2), 492-503*.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**: 185–93.

Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot J-P, Tanay A, et al. 2017. Multiscale 3D genome rewiring during mouse neural development. *Cell* **171**: 557–572.e24.

Cairns J, Freire-Pritchett P, Wingett SW, Várnai C, Dimond A, Plagnol V, Zerbino D, Schoenfelder S, Javierre B-M, Osborne C, et al. 2016. CHiCAGO: Robust detection of dna looping interactions in capture hi-c data. *Genome Biol* **17**: 127.

Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association* **74**: 829–836.

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

Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ. 2015. Condensin-driven remodelling of x chromosome topology during dosage compensation. *Nature* **523**: 240–4.

Cremer T, Cremer M. 2010. Chromosome territories. *Cold Spring Harb Perspect Biol* **2**: a003889.

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

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

Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, et al. 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**: 331–6.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–80.

Djekidel MN, Chen Y, Zhang MQ. 2018a. FIND: DifFerential chromatin interactions detection using a spatial poisson process. *Genome Res*.

Djekidel MN, Chen Y, Zhang MQ. 2018b. FIND: DifFerential chromatin interactions detection using a spatial poisson process. *Genome Res*.

Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI, Zhao K, et al. 2014a. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**: 374–87.

Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI, Zhao K, et al. 2014b. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**: 374–87.

Dozmorov MG, Guthridge JM, Hurst RE, Dozmorov IM. 2010a. A comprehensive and universal method for assessing the performance of differential gene expression analyses. *PLoS One* **5**.

Dozmorov MG, Guthridge JM, Hurst RE, Dozmorov IM. 2010b. A comprehensive and universal method for assessing the performance of differential gene expression analyses. *PLoS One* **5**.

Dudoit S, Yang YH, Callow MJ, Speed TP. 2002a. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica* 111–139.

Dudoit S, Yang YH, Callow MJ, Speed TP. 2002b. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica* 111–139.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016a. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell Syst* **3**: 95–8.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016b. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell Syst* **3**: 95–8.

Edward Y Chen YK Christopher M Tan, Ma'ayan A. 2013. Enrichr: Interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*.

Emma J. Campbell-Smith NWL Nathan M. Holmes, Westbrook RF. 2015. Oxytocin signaling in basolateral and central amygdala nuclei differentially regulates the acquisition, expression, and extinction of context-conditioned fear in rats. *Learning & Memory*.

Fernandez AF, Assenov Y, Martin-Subero JI, Balint B, Siebert R, Taniguchi H, Yamamoto H, Hidalgo M, Tan A-C, Galm O, et al. 2012. A dna methylation fingerprint of 1628 human samples. *Genome Res* **22**: 407–19.

Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, Kraft K, Kempfer R, Jerković I, Chan W-L, et al. 2016. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**: 265–269.

Fraser F J. 2015. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular Systems Biology*.

Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, Barbieri M, Moore BL, Kraemer DCA, Aitken S, et al. 2015. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol* **11**: 852.

Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. 2016. Formation of chromosomal domains by loop extrusion. *Cell Rep* **15**: 2038–49.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**: 58–64.

Ghiselli G, Coffee N, Munnery CE, Koratkar R, Siracusa LD. 2003. The cohesin smc3 is a target the for beta-catenin/tcf4 transactivation pathway. *J Biol Chem* **278**: 20259–67.

GTEx Consortium. 2013. The genotype-tissue expression (gtex) project. *Nat Genet* **45**: 580–5.

Hansen KD, Wu Z, Irizarry RA, Leek JT. 2011. Sequencing technology does not eliminate biological variability. *Nat Biotechnol* **29**: 572–3.

Harewood L, Kishore K, Eldridge MD, Wingett S, Pearson D, Schoenfelder S, Collins VP, Fraser P. 2017. Hi-c as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol* **18**: 125.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol Cell* **38**: 576–89.

Hemberger M, Dean W, Reik W. 2009. Epigenetic dynamics of stem cells and cell lineage commitment: Digging waddington's canal. *Nat Rev Mol Cell Biol* **10**: 526–37.

Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, et al. 2016a. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**: 1454–8.

Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, et al. 2016b. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**: 1454–8.

Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. 2012a. HiCNorm: Removing biases in hi-c data via poisson regression. *Bioinformatics* **28**: 3131–3.

Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. 2012b. HiCNorm: Removing biases in hi-c data via poisson regression. *Bioinformatics* **28**: 3131–3.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012a. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nat Methods* **9**: 999–1003.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012b. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nat Methods* **9**: 999–1003.

Jackson DA, Pombo A. 1998. Replicon clusters are stable units of chromosome structure: Evidence that nuclear organization contributes to the efficient activation and propagation of s phase in human cells. *J Cell Biol* **140**: 1285–95.

Jessica Zuin MIJA van der R Jesse R. Dixon. 2014. Cohesin and ctcf shape the 3D genome differently. *Proceedings of the National Academy of Sciences*.

Jhunjhunwala S, Zelm MC van, Peak MM, Murre C. 2009. Chromatin architecture and the generation of antigen receptor diversity. *Cell* **138**: 435–48.

Ji X, Dadon DB, Powell BE, Fan ZP, Borges-Rivera D, Shachar S, Weintraub AS, Hnisz D, Pegoraro G, Lee TI, et al. 2016a. 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* **18**: 262–75.

Ji X, Dadon DB, Powell BE, Fan ZP, Borges-Rivera D, Shachar S, Weintraub AS, Hnisz D, Pegoraro G, Lee TI, et al. 2016b. 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* **18**: 262–75.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B. 2013a. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**: 290–4.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B. 2013b. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**: 290–4.

John C Stansfield MGD Kellen G Cresswell. 2019. MultiHiCcompare: Joint normalization and comparative analysis of complex hi-c experiments. *Bioinformatics*.

John C. Stansfield TN Duc Tran. 2019. R tutorial: Detection of differentially interacting chromatin regions from multiple hi-c datasets. *Current Protocols in Bioinformatics*.

Kal AJ, Zonneveld AJ van, Benes V, Berg M van den, Koerkamp MG, Albermann K, Strack N, Ruijter JM, Richter A, Dujon B, et al. 1999. Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol Biol Cell* **10**: 1859–72.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AMM, Pletikos M, Meyer KA, Sedmak G, et al. 2011. Spatio-temporal transcriptome of the human brain. *Nature* **478**: 483–9.

Kellen G Cresswell VOPGD John C Stansfield. 2019. SpectralTAD: An r package for defining a hierarchy of topologically associated domains using spectral clustering. *Biorxiv*.

Knight PA, Ruiz D. 2012a. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* drs019.

Knight PA, Ruiz D. 2012b. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* drs019.

Laat W de, Grosveld F. 2003. Spatial organization of gene expression: The active chromatin hub. *Chromosome Res* **11**: 447–59.

Lareau CA, Aryee MJ. 2017. Diffloop: A computational framework for identifying and analyzing differential dna loops from sequencing data. *Bioinformatics*.

Lareau CA, Aryee MJ. 2018. Diffloop: A computational framework for identifying and analyzing differential dna loops from sequencing data. *Bioinformatics* **34**: 672–674.

Lee M, Sheskier M, Farokhnia M, Feng N, Marenco S, Lipska B, Leggio L. 2018. Oxytocin receptor mRNA expression in dorsolateral prefrontal cortex in major psychiatric disorders: A human post-mortem study. *Psychoneuroendocrinology* **96**: 143–147. http://www.sciencedirect.com/science/article/pii/S030645301830132X.

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**: 84–98.

Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009a. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–93.

Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009b. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–93.

Liu X, Yu X, Zack DJ, Zhu H, Qian J. 2008. TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics* **9**: 271.

Lu J, Tomfohr JK, Kepler TB. 2005. Identifying differential expression in multiple sage libraries: An overdispersed log-linear model approach. *BMC Bioinformatics* **6**: 165.

Lun ATL, Smyth GK. 2015a. DiffHic: A bioconductor package to detect differential genomic interactions in hi-c data. *BMC Bioinformatics* **16**: 258.

Lun ATL, Smyth GK. 2015b. DiffHic: A bioconductor package to detect differential genomic interactions in hi-c data. *BMC Bioinformatics* **16**: 258.

Lun ATL, Smyth GK. 2017. No counts, no variance: Allowing for loss of degrees of freedom when assessing biological variability from rna-seq data. *Stat Appl Genet Mol Biol* **16**: 83–93.

Lupiáñez DG, Spielmann M, Mundlos S. 2016a. Breaking tads: How alterations of chromatin domains result in disease. *Trends Genet* **32**: 225–37.

Lupiáñez DG, Spielmann M, Mundlos S. 2016b. Breaking tads: How alterations of chromatin domains result in disease. *Trends Genet* **32**: 225–37.

Ma H, Samarabandu J, Devdhar RS, Acharya R, Cheng PC, Meng C, Berezney R. 1998. Spatial and temporal dynamics of dna replication sites in mammalian cells. *J Cell Biol* **143**: 1415–25.

Marques M WM Hernández RP. 2015. Analysis of changes to mRNA levels and ctcf occupancy upon tfii-i knockdown. *Genom Data*.

McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Res* **40**: 4288–97.

Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, Zhu M, Wu J, Shi X, Taing L, et al. 2017. Cistrome data browser: A data portal for chip-seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res* **45**: D658–D662.

Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, et al. 2015. Mapping long-range promoter contacts in human cells with high-resolution capture hi-c. *Nat Genet* **47**: 598–606.

Miller EK CJ(. 2001. “An integrative theory of prefrontal cortex function”. *Annu Rev Neurosci 24: 167–202*.

Mora A, Sandve GK, Gabrielsen OS, Eskeland R. 2016. In the loop: Promoter-enhancer interactions and bioinformatics. *Brief Bioinform* **17**: 980–995.

Natalie Sauerwald CK Akshat Singhal. 2018. Analysis of the structural variability of topologically associated domains as revealed by hi-c. *Biorxiv*.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, Berkum NL van, Meisig J, Sedat J, et al. 2012. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature* **485**: 381–5.

Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell JA, Lopes S, Reik W, et al. 2004. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* **36**: 1065–71.

O'Sullivan JM, Hendy MD, Pichugina T, Wake GC, Langowski J. a. The statistical-mechanics of chromosome conformation capture. *Nucleus* **4**: 390–8.

O'Sullivan JM, Hendy MD, Pichugina T, Wake GC, Langowski J. b. The statistical-mechanics of chromosome conformation capture. *Nucleus* **4**: 390–8.

Papantonis A, Cook PR. 2013. Transcription factories: Genome organization and gene regulation. *Chem Rev* **113**: 8683–705.

Paulsen J, Sandve GK, Gundersen S, Lien TG, Trengereid K, Hovig E. 2014. HiBrowse: Multi-purpose statistical analysis of genome-wide chromatin 3D organization. *Bioinformatics* **30**: 1620–2.

Phillips JE, Corces VG. 2009. CTCF: Master weaver of the genome. *Cell* **137**: 1194–211.

Phillips-Cremins JE, Corces VG. 2013a. Chromatin insulators: Linking genome organization to cellular function. *Mol Cell* **50**: 461–74.

Phillips-Cremins JE, Corces VG. 2013b. Chromatin insulators: Linking genome organization to cellular function. *Mol Cell* **50**: 461–74.

Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK. 2016. ROBUST hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann Appl Stat* **10**: 946–963.

Popova T, Mennerich D, Weith A, Quast K. 2008. Effect of rna quality on transcript intensity levels in microarray analysis of human post-mortem brain tissues. *BMC Genomics* **9**: 91.

Rao SSP, Huang S-C, Glenn St Hilaire B, Engreitz JM, Perez EM, Kieffer-Kwon K-R, Sanborn AL, Johnstone SE, Bascom GD, Bochkov ID, et al. 2017. Cohesin loss eliminates all loop domains. *Cell* **171**: 305–320.e24. http://dx.doi.org/10.1016/j.cell.2017.09.026.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014a. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665–80.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014b. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665–80.

Rickman DS, Soong TD, Moss B, Mosquera JM, Dlabal J, Terry S, MacDonald TY, Tripodi J, Bunting K, Najfeld V, et al. 2012a. Oncogene-mediated alterations in chromatin conformation. *Proc Natl Acad Sci U S A* **109**: 9083–8.

Rickman DS, Soong TD, Moss B, Mosquera JM, Dlabal J, Terry S, MacDonald TY, Tripodi J, Bunting K, Najfeld V, et al. 2012b. Oncogene-mediated alterations in chromatin conformation. *Proc Natl Acad Sci U S A* **109**: 9083–8.

Robinson MD, McCarthy DJ, Smyth GK. 2010. EdgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–40.

Robinson MD, Smyth GK. 2007. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**: 2881–7.

Robinson MD, Smyth GK. 2008. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics* **9**: 321–32.

Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J, et al. 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* **112**: E6456–65.

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

Sartor MA, Tomlinson CR, Wesselkamper SC, Sivaganesan S, Leikauf GD, Medvedovic M. 2006. Intensity-based hierarchical bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics* **7**: 538.

Scheinin I, Sie D, Bengtsson H, Wiel MA van de, Olshen AB, Thuijl HF van, Essen HF van, Eijk PP, Rustenburg F, Meijer GA, et al. 2014. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res* **24**: 2022–32.

Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL, et al. 2016. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep* **17**: 2042–2059.

Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, Andrews S, Kurukuti S, Mitchell JA, Umlauf D, Dimitrova DS, et al. 2010. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**: 53–61.

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, Heard E, Dekker J, Barillot E. 2015. HiC-pro: An optimized and flexible pipeline for hi-c data processing. *Genome Biol* **16**: 259.

Sexton T, Cavalli G. 2015. The role of chromosome domains in shaping the functional genome. *Cell* **160**: 1049–59.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell* **148**: 458–72.

Shao Z, Zhang Y, Yuan G-C, Orkin SH, Waxman DJ. 2012. MAnorm: A robust model for quantitative comparison of chip-seq data sets. *Genome Biol* **13**: R16.

Shavit Y, Lio' P. 2014a. Combining a wavelet change point and the bayes factor for analysing chromosomal interaction data. *Mol Biosyst* **10**: 1576–85.

Shavit Y, Lio' P. 2014b. Combining a wavelet change point and the bayes factor for analysing chromosomal interaction data. *Mol Biosyst* **10**: 1576–85.

Shipony Z, Mukamel Z, Cohen NM, Landan G, Chomsky E, Zeliger SR, Fried YC, Ainbinder E, Friedman N, Tanay A. 2014. Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* **513**: 115–9.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**: Article3.

Stansfield JC, Cresswell KG, Vladimirov VI, Dozmorov MG. 2018. HiCcompare: An r-package for joint normalization and comparison of hi-c datasets. *BMC Bioinformatics* **19**: 279.

Stouffer SA. 1949. *Adjustment during army life*. Princeton University Press.

Strand AD, Aragaki AK, Baquet ZC, Hodges A, Cunningham P, Holmans P, Jones KR, Jones L, Kooperberg C, Olson JM. 2007. Conservation of regional gene expression in mouse and human brain. *PLoS Genet* **3**: e59.

Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, Ettwiller L, Spitz F. 2014. Functional and topological characteristics of mammalian regulatory domains. *Genome Res* **24**: 390–400.

Taberlay PC, Achinger-Kawecka J, Lun ATL, Buske FA, Sabir K, Gould CM, Zotenko E, Bert SA, Giles KA, Bauer DC, et al. 2016a. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res* **26**: 719–31.

Taberlay PC, Achinger-Kawecka J, Lun ATL, Buske FA, Sabir K, Gould CM, Zotenko E, Bert SA, Giles KA, Bauer DC, et al. 2016b. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res* **26**: 719–31.

Tjur T. 1998. Nonlinear regression, quasi likelihood, and overdispersion in generalized linear models. *The American Statistician* **52**: 222–227. http://www.jstor.org/stable/2685928.

Valton A-L, Dekker J. 2016. TAD disruption as oncogenic driver. *Curr Opin Genet Dev* **36**: 34–40.

Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S. 2015. Comparative hi-c reveals that ctcf underlies evolution of chromosomal domain architecture. *Cell Rep* **10**: 1297–309.

Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, Andrews S. 2015. HiCUP: Pipeline for mapping and processing hi-c data. *F1000Res* **4**: 1310.

Witten DM, Noble WS. 2012. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Res* **40**: 3849–55.

Won H, Torre-Ubieta L de la, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D, et al. 2016. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**: 523–527.

Yaffe E, Tanay A. 2011a. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**: 1059–65.

Yaffe E, Tanay A. 2011b. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**: 1059–65.

Yao B S. 2017. Oxytocin signaling in the medial amygdala is required for sex discrimination of social cues. *eLife, 6, e31373*.

Yu L, Gulati P, Fernandez S, Pennell M, Kirschner L, Jarjoura D. 2011. Fully moderated t-statistic for small sample size gene expression arrays. *Stat Appl Genet Mol Biol* **10**.

Zink D, Fischer AH, Nickerson JA. 2004. Nuclear structure in cancer cells. *Nat Rev Cancer* **4**: 677–87.