

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Analytic Methods for Next-Generation Sequencing Studies of Chromatin Structure and 3D Organization

Permalink

<https://escholarship.org/uc/item/2xn2z1td>

Author

Capurso, Daniel

Publication Date

2015

Peer reviewed|Thesis/dissertation

Analytic Methods for Next-Generation Sequencing Studies of
Chromatin Structure and 3D Organization

by

Daniel Capurso

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Copyright 2015
by
Daniel Capurso

Acknowledgments

This dissertation is the culmination of years of inspiration, encouragement, and support from many people in my life. I am grateful to all those who have been willing to take a chance on me and invest time and resources in my scientific development and to all those who have had my back through obstacles along the way.

My former mentors Drs. Anindya Bagchi and Alea Mills of Cold Spring Harbor Laboratory inspired my interest in genomics. Though I was only a high school student, Dr. Bagchi strongly encouraged my contributing to the intellectual development of the research, rather than just performing assistant work. This was a formative experience in my early life and an opportunity for which I am very grateful.

Many instructors and scientists further cultivated my interest in genomics when I was undergraduate student, including Drs. Rima McLeod, Ruth Lehmann, Douglas Bishop, Wei-Jen Tang, Juan Martinez, and Ilaria Rebay. My former mentors Drs. Dara Torgerson and Dan Nicolae of the University of Chicago provided great guidance in my transition from experimental molecular genetics to bioinformatics.

At UCSF, Drs. Deanna Kroetz and Nadav Ahituv provided outstanding leadership and support in the Pharmaceutical Sciences and Pharmacogenomics (PSPG) Graduate Program, and Rebecca Brown was also very supportive. I thank my Qualifying Exam Committee members – Drs. Nadav Ahituv, Deanna Kroetz, Pui Kwok, and John Witte — and my Thesis Committee members – Drs. Nadav Ahituv and Jun Song — for valuable advice and feedback on this research. I also thank Drs. Hao Xiong and Henrik Bengtsson for sharing some of their knowledge of computer science with me. I thank Drs. Jim Haber and Cheng-Sheng Lee of Brandeis University for a stimulating collaboration.

My career as a graduate student was deeply fulfilling in large part due to the mentorship of Dr. Mark Segal. Mark contributed exceptional creativity in approaching new problems, as well as a very effective mentorship style that aptly balanced flexibility and intellectual freedom with structure and guidance. Mark carries himself with the utmost decency and humility, despite his accolades. For this reason, I consider Mark not only a scientific role model, but also a personal one.

I thank my classmates in the PSPG Graduate Program for maintaining a supportive, collaborative, and fun dynamic. I appreciate the friendships that sustained me during graduate school, especially those of Chelsea, Jimmy, and Keiko.

I am grateful to my parents Mary and Dan for the sacrifices they made to provide for me, which enabled my opportunities and without which this journey might not have been possible. I thank my sister Jess for her enduring support. Though she is my junior, I often look up to her, especially for her strength, insight, and emotional maturity.

Abstract

Beyond linear sequence, higher order structure of the genome influences gene regulation and has been implicated in disease. Chromatin structure is the degree of chromatin compaction at genomic loci. Chromatin organization is the spatial, three-dimensional (3D) positioning of chromatin. Here, we adapt and apply methods for next-generation sequencing analyses of chromatin structure and organization based on chromatin immunoprecipitation-sequencing (ChIP-seq) and genome-wide chromosome conformation capture (Hi-C), respectively. First, we built on a previous study that sought to classify nucleosomes containing either H2A.Z or H2A/H4 arginine 3 symmetric dimethylation (H2A/H4R3me2s) from human ChIP-seq data. We hypothesized that appropriate data preprocessing – deduplication, normalization for sequencing depth, and position-finding – in conjunction with advanced algorithms for feature selection (Discriminatory Motif Feature Selection) and classification (Random Forest) would improve performance. We achieved dramatically improved classification accuracy and identified a significant and biologically meaningful DNA motif associated with H2A/H4R3me2s: “TCCATT”, which is part of the consensus sequence of satellite II and III DNA. Second, we tested our hypothesis that there are advantages to assessing the 3D co-localization of functional annotations (e.g., centromeres) using 3D genome reconstructions from Hi-C contact data because they enable detection of multi-level interactions (assessments using contact data are inherently limited to detecting strictly pairwise interactions). We found significant 3D co-localization of sets of genes with developmentally regulated expression in *Plasmodium falciparum* with 3D reconstruction-based assessment but not with contact-based assessment. Further, we developed a method for 3D reconstruction-based assessment that avoids the data dichotomization of previous approaches. Third, we tested our hypothesis that analyzing ChIP-seq data in combination with

3D reconstructions could identify functional 3D hotspots. We separately overlaid a *Saccharomyces cerevisiae* 3D genome reconstruction with three ChIP-seq inputs and contrasted two algorithms for identifying regions in 3-space — 3D hotspots – for which mean ChIP-seq peak height is significantly elevated: k -Nearest Neighbor (k -NN) regression and the Patient Rule Induction Method (PRIM). For each ChIP-seq input, both algorithms identified significant, corresponding and biologically meaningful 3D hotspots containing distal genomic regions. Our research demonstrates that applying appropriate data preprocessing and advanced supervised learning algorithms improves the interpretability of next-generation sequencing studies of chromatin structure and organization.

Table of Contents

Copyright	ii
Acknowledgments	iii
Abstract	v
Table of Contents	vii
List of Tables	ix
List of Figures	xi
Chapter 1: Background	1
Chapter 2: Preprocessing ChIP-seq data and applying advanced feature selection and classification algorithms to discriminate between histone modifications	11
2.1 Citation	11
2.2 Abstract	11
2.3 Background	12
2.4 Results	16
2.5 Discussion	28
2.6 Conclusions	32
2.7 Methods	33
2.8 Acknowledgments	37

Chapter 3: Distance-based assessment of the localization of functional annotations in 3D genome reconstructions	38
3.1 Citation	38
3.2 Abstract	38
3.3 Background	39
3.4 Results	41
3.5 Discussion	52
3.6 Conclusions	54
3.7 Methods	55
3.8 Acknowledgments	58
Chapter 4: Identifying hotspots in functional genomic data superposed on 3D chromatin configuration reconstructions	59
4.1 Citation	59
4.2 Abstract	59
4.3 Background	60
4.4 Results and Discussion	63
4.5 Conclusions	88
4.6 Methods	90
4.7 Acknowledgments	94
Chapter 5: Discussion	95
References	99

List of Tables

Chapter 2

Table 2.1: Number of sequence reads (or stable nucleosomes) for histone methylations and H2A.Z at each data pre-processing step	18
Table 2.2: Number of sequence reads for histone acetylations at each data pre-processing step	20
Table 2.3: Percentage of sequence reads at each data pre-processing step that contain the motif “TCCATT”	25
Table 2.4: Satellite II and III DNA consensus sequences	26

Chapter 3

Table 3.1: Assessment of the 3D localization of functional annotations in <i>P. falciparum</i> Ring Stage	44
Table 3.2: Assessment of the 3D localization of functional annotations in <i>S. cerevisiae</i>	46
Table 3.3: Comparison of resampling schemes for distance-based assessment of the localization of functional annotations in <i>P. falciparum</i> Ring Stage	48
Table 3.4: Comparison of resampling schemes for distance-based assessment of the localization of functional annotations in <i>S. cerevisiae</i>	49

Chapter 4

Table 4.1: The top 10 Swi6 3D hotspots from k -NN ($k=50$) regression	68
Table 4.2: The top 10 Swi6 3D hotspots from PRIM ($min_beads=25$)	70
Table 4.3: The top 10 Pol2Ser5p 3D hotspots from k -NN ($k=75$) regression	73
Table 4.4: The top 10 Pol2Ser5p 3D hotspots from PRIM ($min_beads=75$)	75
Table 4.5: The top 10 Tup1 3D hotspots from k -NN ($k=50$) regression	78
Table 4.6: The top 10 Tup1 3D hotspots from PRIM ($min_beads=75$)	80
Table 4.7: Top box-ranks of beads from the original Swi6 PRIM 3D hotspot (highlighted in the manuscript) when PRIM ($min_beads = 25$) is applied to rotated 3D reconstructions	83
Table 4.8: Top box-ranks of beads from the original Pol2Ser5p PRIM 3D hotspot (highlighted in the manuscript) when PRIM ($min_beads = 75$) is applied to rotated 3D reconstructions	84
Table 4.9: Top box-ranks of beads from the original Tup1 PRIM 3D hotspot (highlighted in the manuscript) when PRIM ($min_beads = 75$) is applied to rotated 3D reconstructions	84
Table 4.10: Stability of downstream biological findings over settings of k or min_beads	88

List of Figures

Chapter 2

- Figure 2.1: Classifying stable nucleosomes containing H2A/H4R3me2s
or H2A.Z using histone modification features 22
- Figure 2.2: Classifying stable nucleosomes containing H2A/H4R3me2s
or H2A.Z using DNA sequence features 24
- Figure 2.3: Relationship between stables nucleosomes containing histone
modifications and satellite II and III DNA sequences 28

Chapter 3

- Figure 3.1: 3D genome reconstructions 42
- Figure 3.2: Affinity propagation clustering applied to 3D telomere
coordinates for *P. falciparum* Ring Stage 50
- Figure 3.3: Affinity propagation clustering applied to 3D telomere
coordinates for *S. cerevisiae* (HindIII) 51

Chapter 4

- Figure 4.1: ChIP-seq peak height superposed on the 3D chromatin
configuration reconstruction 64
- Figure 4.2: Genomic regions comprising the 1st-ranked Swi6 3D hotspot
from k -NN ($k=50$) regression 69
- Figure 4.3: Genomic regions comprising the 1st-ranked Swi6 3D hotspot
(chr 7;8;16) from PRIM ($min_beads=25$) 71

Figure 4.4: Genomic regions comprising the 1 st -ranked Pol2Ser5p 3D hotspot (chr 10;10;10;13) from k -NN ($k=75$) regression	74
Figure 4.5: Genomic regions comprising the 1 st -ranked Pol2Ser5p 3D hotspot from PRIM ($min_beads=75$)	76
Figure 4.6: Genomic regions comprising the 1 st ranked Tup1 3D hotspot (chr 4;4) from k -NN ($k=50$) regression	79
Figure 4.7: Genomic regions comprising the 4 th ranked Tup1 3D hotspot from PRIM ($min_beads=75$).	81
Figure 4.8: Parameter tuning for k -NN regression and PRIM	87

Chapter 1: Background

Great progress has been made in developing and applying next-generation DNA sequencing technologies (reviewed in¹). This technological revolution has fostered the high-throughput identification of transcribed regions (with RNA-sequencing²), non-coding regulatory regions (e.g., The Encyclopedia of DNA Elements (ENCODE) Project^{3,4}), and population variation (e.g., The 1000 Genomes Project^{5,6}). Alongside these technological and experimental advances, bioinformatics methodologies for predicting and interpreting functional DNA sequences have become increasingly refined. Sequence-based prediction methodologies have been advanced for identifying promoters⁷, exons^{8,9}, splice sites¹⁰, transcription factor binding sites^{11,12}, developmental enhancers¹³, and microRNAs¹⁴.

Considerably less progress has been made in understanding higher order structure of the genome, which also influences genome function and regulation. DNA in the nucleus does not exist as a naked linear string, but rather is compacted as chromatin — the basic unit of which is the nucleosome: ~147 base pairs of DNA wrapped around an octamer of histone proteins. Chemical modifications to DNA and histone tails (referred to collectively as the epigenome) influence *chromatin structure*: the degree of chromatin compaction at each genomic locus and, thus, accessibility to regulatory binding proteins¹⁵. *Chromatin organization* (also called nuclear architecture) is the three-dimensional (3D) spatial positioning of chromatin in the nucleus, which influences, for example, transcriptional activity¹⁶. In this dissertation, we adapt and apply analytic methods to improve the downstream biological interpretability of next-generation sequencing-based studies of chromatin structure and organization.

The compacted, generally transcriptionally repressed form of chromatin is called heterochromatin, while the open, transcriptionally accessible form of chromatin is called

euchromatin. It has been found that DNA methylation and certain histone modifications (e.g., histone H3 lysine 9 trimethylation (H3K9me3)) are associated with heterochromatin, while other histone modifications (e.g., histone H3 lysine 4 trimethylation (H3K4me3)) are associated with euchromatin^{17,18}. Further, in some cases, the molecular mechanism whereby a histone modification affects chromatin accessibility has been elucidated, for example: H3K9me3 provides a bind site for the scaffold protein HP1, which cross-links nucleosomes^{17,18}. So far, modifications have been identified on over 60 histone residues¹⁹. That there are so many different histone modifications with differing genomic localization patterns and in some cases different downstream binding proteins has given rise to the “histone code” hypothesis that individual histone modifications may have specialized functions, such as “indexing” classes of genomic elements²⁰.

Chromatin structure can influence gene expression independently of DNA sequence. Some of the earliest evidence of this was the phenomenon termed position-effect variegation (reviewed in²¹), which was first characterized in *Drosophila* and later studied in mouse (reviewed in²²). Specifically, when a transcriptionally active gene that is normally located in euchromatin becomes relocated adjacent to a heterochromatic region by experimentally induced chromosomal rearrangement, the resulting tissue often has a mosaic phenotype: the gene becomes silenced stochastically in some of the cells (where heterochromatin has “spread”²¹ over the gene) yet remains expressed in other cells (where heterochromatin has not spread over the gene). This phenomenon has also revealed the importance of establishing and maintaining chromatin boundaries with insulators (reviewed in²³).

In addition, chromatin structure is one of the molecular mechanisms whereby the environment can influence phenotypes (reviewed in²⁴), as many signal transduction pathways

have downstream effects on the epigenome (reviewed in²⁵). Thus, even monozygotic human twins have differences in chromatin structure later in life²⁶. These chromatin structure differences and *de novo* somatic mutations can result in some differing phenotypes in adult monozygotic twins. Recent genome-wide DNA methylation analyses of cohorts of such monozygotic twin that are discordant for disease have identified differential DNA methylation of, for example: a serotonin transporter gene (SLC6A4) in bipolar disorder²⁷; a complement factor gene (CFI) in ulcerative colitis²⁸; a hippocampal gene (ZBTB20) in major depressive disorder²⁹; and an insulin pathway gene (MALT1) in type II diabetes³⁰.

Beyond these findings from twin studies, there are many other examples of aberrant DNA methylation and histone modification patterns being associated with human disease and of mutations in the genes that regulate chromatin structure being associated with human disease (reviewed in³¹). Chromatin structure differences are also associated with inter-individual variation in drug response (termed pharmacoeigenomics; reviewed in³²). For example, promoter hypermethylation of certain DNA repair genes (and attendant lower gene expression) in human tumors is associated with greater sensitivity to some chemotherapies (which function by inducing apoptosis via DNA damage). Moreover, a study found that promoter hypermethylation of the DNA-repair gene MGMT in the tumors of glioma patients was significantly associated with a clinical response (decrease in tumor size) to treatment with the alkylating agent carmustine³³. Another study found that hypermethylation of the promoter of the helicase gene WRN in the tumors of colorectal cancer patients was associated with significantly higher median survival time to treatment with the topoisomerase inhibitor irinotecan³⁴.

Similarly, chromatin 3D organization is important for genomic regulation, can affect transcription independently of DNA sequence, and has been implicated in human disease.

Microscopic analyses using chromatin stains or FISH probes have revealed that chromatin is not homogeneously distributed in interphase nuclei, but rather is highly organized with some distinct hallmarks; for example, in *S. cerevisiae*: telomeres and heterochromatin tend to localize near the nuclear periphery, centromeres tend to cluster near the spindle pole body, and the chromosome 12 ribosomal DNA (rDNA) repeat region tends to localize near the nucleolus^{35,36}. Some of the molecular mechanisms contributing to this organization have been elucidated, for example: microtubule attachments from the centromere kinetochores to the spindle pole body, and the binding of telomeric proteins with nuclear membrane proteins (e.g., of Sir4 with Esc1)³⁵.

Studies have also provided evidence that chromatin organization can affect gene expression independently of DNA sequence. For example, physically tethering a reporter gene to the nuclear membrane in *S. cerevisiae* (via a recombinant protein with an integral membrane domain and DNA binding domain) resulted in transcriptional repression of the reporter gene³⁷. Chromatin organization has also been implicated in human disease. For example, lamins are nuclear membrane proteins in humans that normally bind heterochromatin, contributing to its localization near the nuclear periphery; a mutation in Lamin A causes human progeria (an accelerated aging disease), resulting from loss of constitutive heterochromatin and aberrant transcription of repeat regions³⁸. Chromatin organization also likely influences the location of DNA breakpoints and gene fusions³⁹, including those that drive certain cancers⁴⁰.

The development of two next-generation sequencing-based protocols has enabled chromatin structure and organization to be inferred at high resolution: chromatin immunoprecipitation sequencing (ChIP-seq)⁴¹ and genome-wide chromosome conformation capture (Hi-C^{42,43}), respectively. ChIP-seq is performed as follows. Cells are treated with formaldehyde to cross-link DNA with the proteins that are bound to it. The DNA (along with the

cross-linked proteins) is then fragmented with micrococcal nuclease (MNase) digestion (or in some cases sonication). An antibody that specifically recognizes a histone modification (or chromosome associated protein) of interest is then used to immunoprecipitate (IP) the chromatin. The DNA and proteins are then reverse cross-linked and the DNA is analyzed with next-generation sequencing. This process enriches the experimental sample for reads coming from DNA fragments associated with the histone modification of interest; however, much of the sample is still background reads⁴⁴. If a control (“mock IP”) sample is prepared – by using a non-specific antibody during the IP step and performing the same protocol — the experimental sample can be normalized to the control sample.

Hi-C is performed as follows. Genomic interactions are captured by cross-linking protein to DNA with formaldehyde (the logic being that if two regions of the genome are physically interacting, then that interaction is mediated by a protein). The DNA is then cut with a restriction enzyme, and re-ligated under dilute conditions to promote intramolecular ligations so that the two interacting pieces of DNA are now juxtaposed in a single circular fragment. The protein is then reverse cross-linked from the DNA and an adaptor is added to the circular DNA fragment, which is then paired-end sequenced. Unlike most paired-end sequencing experiments, in this case, we expect for each mate pair to map to a different part of the genome (corresponding to each of the two genomic regions that were physically interacting). The resulting contact data from Hi-C analysis list two genomic positions – each corresponding to a restriction enzyme site (or bin if the data are binned) – and the number of times they were paired-end sequenced together. The higher this interaction frequency, the smaller the physical distance should be between the two genomic sites. Using differing means for quantifying this relationship a variety of approaches for generating 3D genome reconstructions from the contact data have been

advanced. These include constrained optimization of a multi-dimensional scaling criterion^{43,45}, where the constraints derive from prior biological and biophysical knowledge, e.g., chromatin contiguity and avoidance of steric clashes.

In this dissertation, we adapt and apply analytic methods to improve the downstream biological interpretability of next-generation sequencing studies of chromatin structure (based on ChIP-seq data) and organization (based on 3D genome reconstructions from Hi-C data). In Chapter 2, we build on a previous analysis that applied classification algorithms to attempt to discriminate between histone modifications from a ChIP-seq dataset for 20 histone methylations and the histone variant H2A.Z⁴⁶ and a ChIP-seq dataset for 18 histone acetylations⁴⁷ using as features DNA sequence motifs in the nucleosomal DNA and co-occurrence with other histone modifications. Such a computational undertaking is exciting because it could reveal aspects of the histone code hypothesis, which has been difficult to probe experimentally because of genetic redundancy and enzyme promiscuity⁴⁸. Specifically, Gervais and Gaudreau⁴⁹ attempted to discriminate between nucleosomes containing the histone modification H2A/H4R3me2s (symmetric Arginine 3 dimethylation of histones H2A and/or H4; the antibody recognizes both) or the histone variant H2A.Z. H2A.Z and H2AR3me2s should be mutually exclusive because H2A.Z has a truncation relative to H2A, such that the R3 site is not present. However, the authors only attained modest classification accuracy and with limited downstream biological interpretation⁴⁹.

We identified two potential analytic issues that could have contributed to the modest performance: inadequate preprocessing of the ChIP-seq data (they used raw reads), and the use of less sophisticated algorithms than are available, both for feature selection (they used enumerative feature generation, e.g. all k -mers, which restricts k to being relatively small for

computational reasons) and for classification (they used a C4.5 decision tree, which does not have the performance benefits of an ensemble classifier). Accordingly, we hypothesized that appropriate ChIP-seq data pre-processing in conjunction with more advanced feature selection and classification algorithms could improve performance both in terms of classification accuracy and interpretative yield.

For ChIP-seq data preprocessing we: deduplicated reads to eliminate PCR amplification bias^{50,51}, normalized for unique read number via down-sampling to control for bias from variable sequencing depth⁵¹, and identified stable nucleosomes with significant enrichment over the background using NPS⁴⁴. Though the Barski dataset⁴⁶ is a very rich resource of histone modification ChIP-seq data from a human primary cell line, it does not have a mock IP control sample. However, the fact that we are performing comparative analyses across experimental samples somewhat circumvents the problem of a lack of mock IP control sample (i.e., any systematic biases should be the same in both experimental samples and thus will not contribute to a discriminative signal between the samples). For feature selection, we used Discriminatory Motif Feature Selection (DMFS)⁵², which generates a small set of motifs that discriminate between the two classes *a priori* using a partition of the data (such that a different data subset is used for motif discovery than for classification, with the latter subset further partitioned into a training set and validation set). In contrast to enumerative feature generation, DMFS avoids the generation of extensive noise features— which can degrade classifier performance⁵³ – and allows longer, potentially more informative motifs to be evaluated. For classification, we used Random Forests⁵⁴, which has performance gains as result of averaging over ensembles of classification trees. An additional advantage of Random Forests is that it still emphasizes interpretation by allowing features to be ranked by importance. This is possible because each tree is constructed

from a bootstrap sample of the data, and thus the data points that were not sampled at each tree (so-called “out-of-bag” (OOB) data points) can then be used to evaluate feature importance.

In Chapter 3, we sought to improve the interpretability of methods for assessing the 3D co-localization of functional annotations (e.g., centromeres and long terminal repeats) from 3D genome reconstructions. These analyses make use of *Saccharomyces cerevisiae* and the malaria parasite *Plasmodium falciparum* because these organisms are haploid and because their relatively smaller genomes (compared to human) enable higher resolution Hi-C data and the generation of 3D genome reconstructions (a mammalian 3D genome reconstruction has not yet been generated for computational reasons). Previously, Witten and Noble⁵⁵ assessed functional annotation 3D co-localization in *S. cerevisiae* from the contact data (but not from the 3D genome reconstruction), while Ay et al.⁴⁵ performed such assessments in *P. falciparum* from the 3D genome reconstruction (but not from the contact data).

The first novel contribution we made was to make a side-by-side comparison of contact-based and 3D reconstruction-based assessment of functional annotation 3D localization in each organism because we hypothesized that there are advantages to analyzing the 3D reconstructions: (i) while the contact data is inherently limited to detecting strictly pairwise interactions, the 3D reconstructions enable detection of multi-level interactions; (ii) the 3D location of sites for which there is missing contact data is readily determined from neighboring points in the reconstruction because of chromatin contiguity; and (iii) biological and biophysical constraints about genome organization are imposed (e.g. avoidance of steric clashes). Thus, emergent properties of the 3D reconstructions may reveal significant co-localization of some functional annotations that were not co-localized in the (pairwise) contact data.

Another potential drawback of both previous analyses is that they dichotomized data point pairs as “close” or “far” based on some threshold (e.g., 10%, 20%, or 40% of the nuclear diameter in 3D reconstruction-based assessment) and then tested for enrichment of “close” pairs for each functional annotation. In some cases, this led to significance that varied by dichotomization threshold, and it is not obvious what constitutes a biologically meaningful threshold choice. We hypothesized that reconstruction-based assessment of functional annotation 3D co-localization using a summary statistic (the Median of Pairwise Euclidean Distances (MPED)) could replicate some of the key biological findings without producing threshold-sensitive results. We chose to use the *median* (because of its robustness and resistance properties) of *all* pairwise distances (because this does not require tuning as, for example, would be necessary with k nearest neighbor distances).

In Chapter 4, we hypothesized that analyzing ChIP-seq data in combination with 3D genome reconstructions from Hi-C data could enable the detection of functional nuclear hotspots. First, we separately superposed three ChIP-seq inputs (normalized to a mock IP control) onto a 3D genome reconstruction. Then we adapted, applied, and contrasted two algorithms for identifying regions in 3-space – “3D hotspots” – for which the mean ChIP-seq peak height is significantly elevated: k -Nearest Neighbor (k -NN) regression⁵⁶ and the Patient Rule Induction Method (PRIM)^{57,58}. While a couple of previous studies superposed functional genomic data onto 3D genome reconstructions, they did so solely for visualization⁵⁹ or to assess *global* concordance of the functional genomic data with the 3D genome reconstruction⁴⁵. An advantage of our novel focal analyses is that assessment of the gene membership of so elicited 3D hotspots can then reveal valuable downstream biological information. The ChIP-seq inputs were: Swi6, a transcription factor; RNA polymerase II phosphorylated at Serine 5 (Pol2Ser5p),

the active transcriptional machinery; and Tup1, a repressor. An advantage of k -NN regression is that it is invariant to rotations of the 3D reconstruction (which is coordinate-free). On the other hand, PRIM is arguably more robust to parameter tuning and, though it is not invariant under rotations of the 3D reconstruction, its rotational dependence can be assessed by analyzing (disparate) rotated 3D reconstructions.

Chapter 2: Preprocessing ChIP-seq data and applying advanced feature selection and classification algorithms to discriminate between histone modifications

2.1 Citation

Capurso D, Xiong H, Segal MR (2012). A histone arginine methylation localizes to nucleosomes in satellite II and III DNA sequences in the human genome. *BMC Genomics*, **13**:630.

2.2 Abstract

Background

Applying supervised learning / classification techniques to epigenomic data may reveal properties that differentiate histone modifications. Previous analyses sought to classify nucleosomes containing histone H2A/H4 arginine 3 symmetric dimethylation (H2A/H4R3me2s) or H2A.Z using human CD4⁺ T-cell chromatin immunoprecipitation sequencing (ChIP-seq) data. However, these efforts only achieved modest accuracy with limited biological interpretation. Here, we investigate the impact of using appropriate data pre-processing —deduplication, normalization, and position- (peak-) finding to identify stable nucleosome positions — in conjunction with advanced classification algorithms, notably discriminatory motif feature selection and random forests. Performance assessments are based on accuracy and interpretative yield.

Results

We achieved dramatically improved accuracy using histone modification features (99.0%; previous attempts, 68.3%) and DNA sequence features (94.1%; previous attempts, <60%). Furthermore, the algorithms elicited interpretable features that withstand permutation

testing, including: the histone modifications H4K20me3 and H3K9me3, which are components of heterochromatin; and the motif TCCATT, which is part of the consensus sequence of satellite II and III DNA. Downstream analysis demonstrates that satellite II and III DNA in the human genome is occupied by stable nucleosomes containing H2A/H4R3me2s, H4K20me3, and/or H3K9me3, but not 18 other histone methylations. These results are consistent with the recent biochemical finding that H4R3me2s provides a binding site for the DNA methyltransferase (Dnmt3a) that methylates satellite II and III DNA.

Conclusions

Classification algorithms applied to appropriately pre-processed ChIP-seq data can accurately discriminate between histone modifications. Algorithms that facilitate interpretation, such as discriminatory motif feature selection, have the added potential to impart information about underlying biological mechanism.

2.3 Background

Chromatin compaction is one of the critical factors regulating gene expression. The basic unit of chromatin, the nucleosome, consists of 147 base pairs (bp) of DNA wrapped around an octamer of histone proteins (H2A, H2B, H3, H4). Many histone post-translational modifications contribute to establishing compacted, transcriptionally repressed *heterochromatin* (e.g., histone H3 lysine 9 trimethylation (H3K9me3)) or open, transcriptionally poised *euchromatin* (e.g., H3K4me3)^{15,17}. However, it is currently unknown why so many modifications — on at least 60 histone residues¹⁹ — are necessary^{18,19}. One possibility is that individual modifications have specialized properties, such as “indexing” classes of genomic elements⁶⁰. Nevertheless, such

discriminating properties remain largely unknown, as redundancy and enzyme promiscuity for non-histone targets have limited the amenability of histone modifications to genetic experimentation⁴⁸.

A potential solution to this problem is to apply supervised learning / classification techniques to high-throughput epigenomic data, such as chromatin immunoprecipitation sequencing (ChIP-seq) data, for histone modifications. Encouragingly, these approaches have had success in the related task of predicting the nucleosome occupancy of DNA sequences: they have elicited predictive features with biological (e.g., Rap1 transcription factor binding sites^{61,62}) and biophysical (e.g., GC content, DNA propeller twist^{61,63,64}) interpretations. Nevertheless, attempts to apply classification techniques to histone modifications have been less forthcoming. This is, in part, because such analyses require richer, thus less readily available, datasets, which correspond to many ChIP-seq experiments in the same cell type. As notable exceptions, Barski *et al.*⁴⁶ have generated a ChIP-seq dataset for 20 histone methylations and the histone variant H2A.Z in human CD4⁺ T cells, and Wang *et al.*⁴⁷, of the same research group, have generated a similar dataset for 18 histone acetylations. A recent study by Gervais and Gaudreau⁴⁹ applied classification techniques to histone modifications using these datasets.

In particular, Gervais and Gaudreau⁴⁹ attempted to predict whether a nucleosome contains histone H2A.Z or H2A/H4 arginine 3 symmetric dimethylation (H2A/H4R3me2s; the authors refer to this as just “H2A”, though it is a methylated form⁶⁵). Importantly, these two classes are likely mutually exclusive: H2A.Z lacks the R3 methylation site and localizes near active transcription start sites¹⁵, while H2A/H4R3me2s localizes with repressed heterochromatin¹⁹. The authors⁴⁹ first performed classification with histone modification features (co-localization with 37 other modifications from ChIP-seq) and, then, with DNA sequence

features (frequency of 6-mers in 147bp nucleosome-bound DNA sequences). However, these analyses only achieved modest prediction accuracies of 68.3% and <60%, respectively (here, a trivial classifier would have an accuracy of 50%)⁴⁹. Furthermore, there was limited biological interpretation for histone modification features and no interpretation for DNA sequence features⁴⁹.

A partial explanation for this modest performance may be insufficient data pre-processing. First of all, Gervais and Gaudreau⁴⁹ used *raw*, aligned (25 base pair) ChIP-seq reads, and simply extended these to 147 base pairs to generate what they consider to be nucleosome-bound DNA sequences. However, this approach is problematic. Because ChIP-seq is only a slight enrichment (not a purification) for sequences bound to the protein of interest⁵⁰, it is notoriously noisy. The majority (estimates upward of 90%⁶⁶) of ChIP-seq reads are instead from the background. Therefore, we, and others^{44,50,51}, advocate using position- (peak-) finding algorithms, such as Nucleosome Positioning from Sequencing (NPS)⁴⁴ (see *Methods*), that identify stable nucleosome positions, with statistically significant enrichment over background, prior to analysis. Here, *stable* nucleosomes can be defined as those that are located at roughly the same chromosomal position across a population of cells and can therefore generate a signal peak when ChIP-seq reads are aligned. Such nucleosomes are also referred to as being relatively well *positioned* or *phased*, and there is evidence for their regulatory importance^{15,66}. While using stable nucleosome positions might limit the analysis to a subset of nucleosomes (and thus influence interpretation), we still believe this approach is preferable to using raw, aligned reads — of which only a small minority were likely even bound to the nucleosomes of interest. This approach of using stable nucleosomes was also utilized in a recent study⁶⁷.

Aside from this issue of the handling of signal and background, the approach used in Gervais and Gaudreau⁴⁹ might not adequately control for systematic biases present in ChIP-seq data. Because of PCR amplification bias in ChIP-seq data, it may be advisable to collapse duplicate reads prior to analysis^{50,51}, especially in datasets such as Barski *et al.*⁴⁶ and Wang *et al.*⁴⁷ where sequencing depth is relatively low, such that there is a lower likelihood of sequencing independently-precipitated fragments with the same start site. Even in the case of stable nucleosomes, the positioning is often blurry, with nucleosomes not having precisely the same start site across cells⁶⁸. However, it is important to note that as future datasets begin to have much higher sequencing depth as a result of decreasing sequencing costs, more refined techniques are needed to control PCR bias than simply collapsing duplicate reads. In addition to PCR bias, coverage and the ability to detect peaks vary with sequencing depth, so ChIP-seq experiments need to be normalized for the number of reads prior to analysis⁵¹. Refined normalization approaches are emerging⁶⁹ for ChIP-seq datasets that contain a mock immunoprecipitation (IP) sample; however, for otherwise rich ChIP-seq datasets that lack such a mock IP, including⁴⁶ and⁴⁷, we believe data should still be normalized for the number of reads, in the absence of a more delicate normalization method for this type of data (see Discussion).

Here, we employ appropriate ChIP-seq data pre-processing and sequence-customized, or otherwise advanced, algorithms to investigate their impact on the accuracy and interpretability of classifying nucleosomes containing H2A/H4R3me2s or H2A.Z. For data pre-processing, we perform deduplication, normalization, and position-finding. Further, for DNA sequence-based classification, we utilize the recently developed Discriminatory Motif Feature Selection (DMFS)⁵², which, in addition to achieving impressive accuracy, emphasizes interpretability, unlike so-called “black-box” classifiers. Specifically, DMFS elicits a small set of *a priori*

discriminatory features (motifs) on a subsequently withheld data partition. This eliminates many noise features, which can comprise prediction and interpretation⁵³, and loosens restrictive feature length prescriptions (e.g., 6-mers in⁴⁹), which could otherwise fail to generate key, longer features. For classification based on histone modification features, we utilize an ensemble method, random forests⁵⁴, which have been widely demonstrated to improve on individual classification trees^{54,58}, as were deployed by Gervais and Gaudreau⁴⁹. Finally, we perform extensive downstream analysis. Importantly, in addition to achieving dramatically improved accuracies, our classification algorithms elicit predictive, interpretable features that are consistent with recent biochemical findings⁷⁰.

2.4 Results

We pre-processed the Barski *et al.*⁴⁶ ChIP-seq datasets for 20 histone methylations and the histone variant H2A.Z to reduce bias. The percentage of duplicate reads in each experiment ranged from 2.1% to 25.1% (*median* = 5.6%), suggesting the potential for substantial PCR bias in some of the samples. We therefore collapsed duplicate reads into single reads. Additionally, the number of unique reads in the experiments varied by more than 3-fold, indicating the potential for considerable sequencing depth variation (and thus coverage bias) across the raw samples. We therefore normalized experiments for sequencing depth by down-sampling to the lowest number of unique reads observed (see *Methods*).

Using this filtered data, we identified stable nucleosome positions as signal peaks with statistically significant enrichment over the background by applying NPS⁴⁴ (see *Methods*). This yielded 1845 and 46235 stable nucleosomes containing H2A/H4R3me2s and H2A.Z, respectively (Table 2.1). Next, we down-sampled H2A.Z nucleosomes to match the number of

H2A/H4R3me2s nucleosomes for two reasons. First, this creates a balanced dataset (i.e., where a trivial classifier has an accuracy of 50%) for subsequent classification and yields accuracies directly comparable to those of ⁴⁹ (who performed analogous down-sampling). Indeed, using “class-imbalanced” data can result in classifier that is biased toward the larger class, as is discussed and investigated in ⁷¹; the authors also demonstrate that, in the case of high-dimensional data, down-sampling the larger class is preferable to over-sampling the smaller class. Second, down-sampling emphasizes features associated with H2A/H4R3me2s, which is relatively under-studied compared to H2A.Z. An added benefit of this approach is its reduction of the computational burden. All reported performance results are the mean of (cross-validated or out-of-bag) performance summaries over 10 different random down-samplings of H2A.Z nucleosomes — this, to ensure our balanced approach did not bias the results.

Histone Modification	Number of Sequence Reads				Number of Stable Nucleosomes
	Sequencing <i>Barski et al., 2007</i>	Alignment <i>Barski et al., 2007</i>	Deduplication <i>Present study</i>	Normalization <i>Present study</i>	Position-finding <i>Present study</i>
H2A/H4R3me2s	25,128,493	7,357,597	7,140,521	4,330,278	1,854
H2A.Z	14,641,244	7,536,100	6,726,630	4,330,278	46,235
H2BK5me1	21,230,477	8,942,880	7,938,208	4,330,278	42,165
H3K4me1	37,461,698	11,322,526	9,921,429	4,330,278	74,544
H3K4me2	13,088,174	5,447,902	5,234,477	4,330,278	56,403
H3K4me3	39,872,596	16,845,478	13,344,169	4,330,278	64,453
H3K9me1	16,446,697	9,311,627	8,824,220	4,330,278	32,385
H3K9me2	19,712,420	9,782,127	9,411,727	4,330,278	282
H3K9me3	12,284,114	6,348,997	5,941,216	4,330,278	4,923
H3K27me1	20,481,466	10,047,279	9,705,780	4,330,278	2,132
H3K27me2	20,998,788	9,070,882	8,862,687	4,330,278	458
H3K27me3	28,475,252	8,970,141	8,632,665	4,330,278	689
H3K36me1	12,898,612	8,077,127	7,907,199	4,330,278	410
H3K36me3	30,015,905	13,572,575	12,362,519	4,330,278	14,495
H3K79me1	19,253,958	5,137,886	4,979,854	4,330,278	32,936
H3K79me2	4,341,935	4,712,875	4,448,350	4,330,278	84,571
H3K79me3	21,024,126	5,929,782	4,440,702	4,330,278	72,059
H3R2me1	16,465,425	9,560,224	9,195,984	4,330,278	602
H3R2me2a	14,743,869	6,521,560	5,953,869	4,330,278	571
H4K20me1	20,396,442	11,015,873	9,640,668	4,330,278	70,084
H4K20me3	18,380,292	5,720,089	4,330,278	4,330,278	17,962

Table 2.1: Number of sequence reads (or stable nucleosomes) for histone methylations and H2A.Z at each data pre-processing step.

In the columns labeled “Sequencing” through “Normalization”, cells indicate the number of 25-base pair sequence reads (from Barski et al.⁴⁶) that are retained at each data pre-processing step. In the final column, cells indicate the number of signal peaks that correspond to stable nucleosomes, identified using NPS.

Classification using additional histone modification features

The presence of one type of histone modification in a nucleosome can increase or decrease the likelihood of a second type¹⁷. Therefore, to identify such potential interactions, we attempted to discriminate between stable nucleosomes containing H2A/H4R3me2s or H2A.Z by using the co-localization with 19 remaining histone methylations and 18 histone acetylations (Table 2.2) as features for classification. For each stable nucleosome, we generated an array of length 37 (for 37 feature modifications), where each entry is the number of deduplicated sequence reads for a feature modification that map within the nucleosome boundaries in a strand-specific manner (see *Methods*). The motivation for using deduplicated sequence read counts for scoring overlap with feature modifications is that it results in a richer (i.e., less sparse) feature matrix than scoring binary overlap with stable nucleosomes for the feature modifications. We still use stable nucleosomes, however, for the outcome modifications (H2A/H4R3me2s, H2A.Z) and in downstream analyses.

Histone Modification	Number of Sequence Reads		
	Sequencing <i>Wang et al., 2008</i>	Alignment <i>Wang et al., 2008</i>	Deduplication <i>Present study</i>
H2AK5ac	9,260,603	3,442,542	3,400,544
H2AK9ac	4,228,439	2,070,246	1,882,070
H2BK5ac	7,635,650	3,330,268	3,066,338
H2BK12ac	9,438,261	3,615,226	3,515,350
H2BK20ac	7,868,594	4,083,727	3,929,081
H2BK120ac	7,693,057	3,444,551	3,280,474
H3K4ac	7,255,253	3,546,672	3,438,276
H3K9ac	9,357,424	3,950,661	3,726,987
H3K14ac	8,987,513	3,799,058	3,755,104
H3K18ac	9,227,062	4,249,604	4,046,345
H3K23ac	7,313,742	2,527,421	2,510,612
H3K27ac	8,529,409	3,433,165	3,198,818
H3K36ac	8,934,172	4,374,235	4,196,023
H4K5ac	8,829,494	4,118,574	4,020,280
H4K8ac	8,350,731	4,278,905	4,176,246
H4K12ac	6,641,441	3,677,187	3,602,609
H4K16ac	19,471,237	7,059,753	6,921,635
H4K91ac	5,087,302	3,191,156	3,016,564

Table 2.2: Number of sequence reads for histone acetylations at each preprocessing step. Cells indicate the number of 25-base pair sequence reads (from Wang et al.⁴⁷) that are retained at each data pre-processing step.

We attained highly accurate random forest (see *Methods*) prediction performance using histone modification features, with an accuracy of $99.0\% \pm 0.1\%$ and an area under the Receiver Operating Characteristic curve (auROC) of 0.999 ± 0.0002 (Figure 2.1a). This is a substantial improvement over the corresponding accuracy of 68.3% that Gervais and Gaudreau⁴⁹ report. To determine which features were “driving” the classification, we evaluated random forest feature

importance by mean decrease in Gini index (MDG; Figure 2.1b; see *Methods*). Several features ranked prominently and withstood estimation of statistical significance by permutation testing (see *Methods*): H4K20me3, H3K9me3, H3R2me2a, H3K36me3, H3K18ac, H3K9me2, and H3K27ac had a permutation $p < 1e-05$ (Bonferroni-adjusted $p < 3.7e-04$; Figure 2.1b). The remaining histone modification features were not significant.

To further explore how these features relate to H2A/H4R3me2s, we built a single classification tree (Figure 2.1c)⁷², which, compared to the random forest ensemble of trees, may more readily reveal interpretable rules, albeit at the cost of decreased classification accuracy. Consistent with the random forest feature importance ranking, the feature that best separated the data in the single tree is H4K20me3 (Figure 2.1c). Indeed, 1737 out of 1854 stable nucleosomes containing H2A/H4R3me2s were classified at the first split, based on overlapping with greater than two deduplicated, H4K20me3 sequence reads (with a misclassification rate of only 1.67%). Three of the four remaining splits were also based on features that had significant random forest feature importance (H3K18ac, H3K27ac, and H3R2me2a; H2BK5me1 did not have a significant random forest feature importance, yet was the basis for the second split). H3K9me3, which had the second highest random forest feature importance, was not the basis for a split in the single tree; however, this may occur if, for example, the stable H2A/H4R3me2s nucleosomes that overlap with H3K9me3 are a subset of those that overlap with H4K20me3 (and so they are already classified at the first split).

Encouragingly, the top two modifications by random forest feature importance, H4K20me3 and H3K9me3, are more frequent in stable nucleosomes containing H2A/H4R3me2s than those containing H2A.Z (Figure 2.1b). Because H4K20me3 and H3K9me3 have been

shown to contribute to the formation of heterochromatin^{15,17} – which is where H2A/H4R3me2s localizes — this initial finding supports the biological relevance of our classifier.

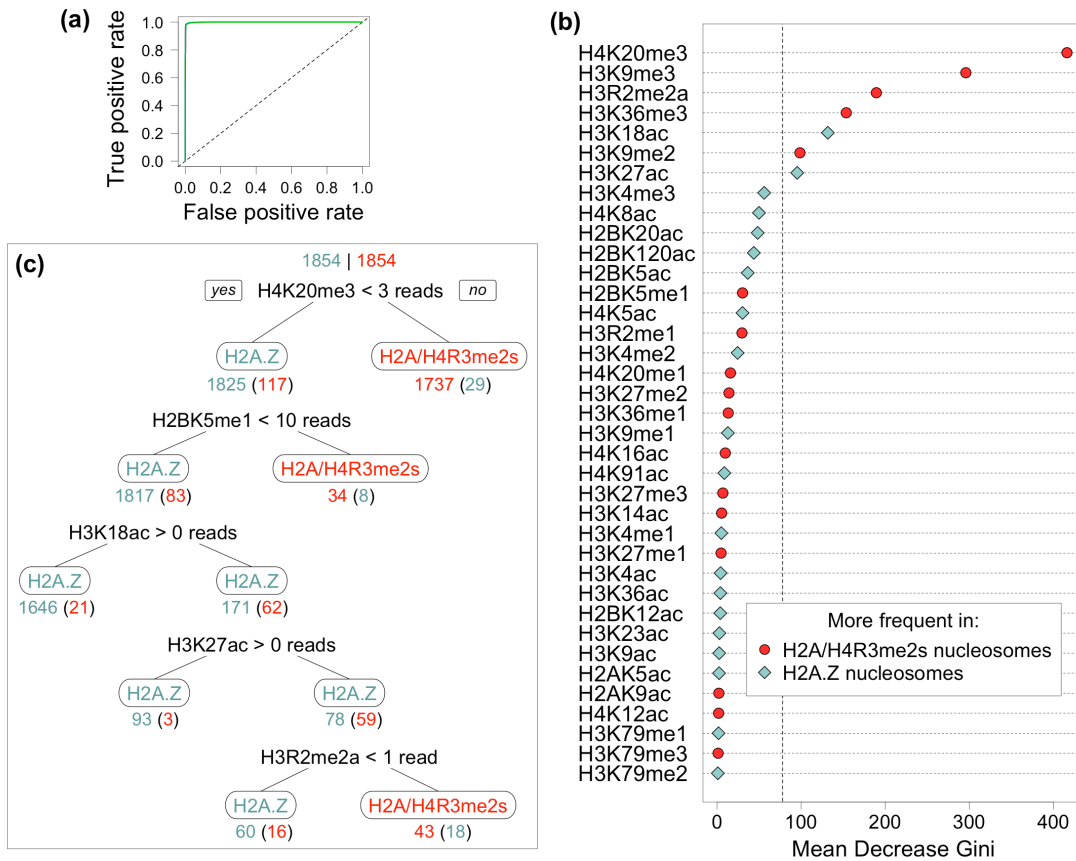


Figure 2.1: Classifying stable nucleosomes containing H2A/H4R3me2s or H2A.Z using histone modification features

(a) Receiver Operating Characteristic (ROC) curve, demonstrating classifier performance. (b) Random forest feature importance by mean decrease in Gini index. Features have a higher frequency in H2A/H4R3me2s nucleosomes (red) or H2A.Z nucleosomes (blue). The dashed, vertical line shows the estimated (permutation-based) significance threshold after multiple testing correction. (c) A classification tree with splits (no borders) and leaves (borders), below which is the number of nucleosomes classified correctly and, in parentheses, incorrectly at that stage. Leaves show the predicted class labels of nucleosomes partitioned there. Splits show the condition that best separates the data. Branch labels indicate the directions in which the split condition is true (“yes”) and false (“no”).

Classification using DNA sequence features

DNA sequence likely influences the genome-wide distribution of histone modifications, as sequence-specific transcription factors and microRNAs can bind and recruit histone-modifying enzymes⁷³. Thus, we used DNA sequence motifs as features for classifying H2A/H4R3me2s and H2A.Z nucleosomes for two reasons: first, to identify such potential targeting sequences, and second, to identify classes of genomic elements that the histone modification potentially regulates. Using DMFS⁵², we identified <300 *a priori* discriminatory motifs with lengths between 5 and 10 bp from a subsequently withheld partition of the data (see *Methods*).

As above, we attained highly accurate random forest prediction performance using DNA sequence features (discriminatory motifs), with an accuracy of 94.1% \pm 0.3% (auROC = 0.968 \pm 0.001; Figure 2.2a). This is a dramatic improvement over the corresponding accuracy of <60% that Gervais and Gaudreau⁴⁹ report. We next evaluated random forest feature importance by MDG (see *Methods*). The top 20 features (Figure 2.2b), all of which occur more frequently in DNA corresponding to stable H2A/H4R3me2s nucleosome positions, withstand estimation of statistical significance by permutation testing, with permutation $p < 1e-05$ (Bonferroni-adjusted $p < 2.7e-03$). Interestingly, 12 of these 20 sequence features contain the motif TCCATT (Figure 2.2b). We therefore analyzed the frequency distribution of the number of occurrences of this motif in the DNA sequences corresponding to stable nucleosome positions (Figure 2.2c, Table 2.3). Indeed, while the motif TCCATT is present in only ~7% of stable H2A.Z nucleosomal DNA sequences ($max = 3$ occurrences per sequence), it is present in ~72% of stable H2A/H4R3me2s nucleosomal DNA sequences ($max = 23$ occurrences per sequence; $median = 7$; Figure 2.2c). That this 6-mer occurs so abundantly in many of the stable H2A/H4R3me2s

nucleosomal DNA sequences is suggestive of it being a repetitive element, or component thereof – an observation we explore in downstream analysis.

For thoroughness, however, we first performed a combined classification that utilized histone modification features *and* DNA sequence features. This resulted in a classification accuracy of $98.6\% \pm 0.1\%$ ($\text{auROC} = 0.999 \pm 0.0002$). Feature importance analysis by MDG yielded many of the same top features as in the separate classifications, namely: H4K20me3, H3K9me3, H3R2me2a, H3K36me3, and sequences containing the motif TCCATT.

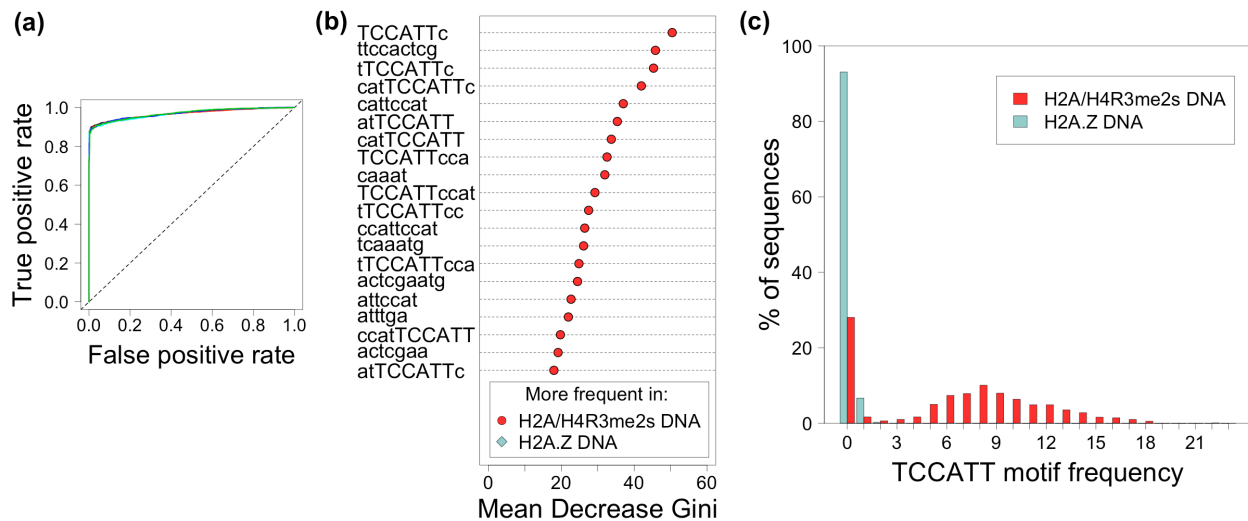


Figure 2.2: Classifying stable nucleosomes containing H2A/H4R3me2s or H2A.Z using DNA sequence features

(a) Receiver Operating Characteristic (ROC) curve, demonstrating classifier performance. (b) Random forest feature importance by mean decrease in Gini index. Features have a higher frequency in H2A/H4R3me2s nucleosomal DNA (*red*) or H2A.Z nucleosomal DNA (*blue*). (c) Frequency histogram of the number of occurrences of the motif TCCATT in H2A/H4R3me2s nucleosomal DNA (*red*) or H2A.Z nucleosomal DNA (*blue*).

Histone Modification	Sequence reads (%) containing the motif “TCCATT”				
	Sequencing <i>Barski et al., 2007</i>	Alignment <i>Barski et al., 2007</i>	Deduplication <i>Present study</i>	Normalization <i>Present study</i>	Position-finding <i>Present study</i>
H2A/H4R3me2s	3.70	2.96	2.52	2.51	56.79
H2A.Z	1.11	1.38	1.42	1.42	0.91
H3K9me3	4.41	2.48	1.95	1.95	9.24
H4K20me3	20.44	11.34	4.59	4.59	10.53

Table 2.3: Percentage of sequence reads at each data pre-processing step that contain the motif “TCCATT”

Cells indicate the percentage of 25-base pair sequence reads that contain at least one occurrence of the motif “TCCATT”. The column labeled “Position-finding” uses the 25-base pair sequence reads that contribute to the signal peaks of stable nucleosomes, identified using NPS.

Downstream feature analysis

Having elicited important, predictive features (particularly H4K20me3, H3K9me3, and the sequence motif TCCATT), we pursued downstream analysis in an attempt to determine how they relate functionally to H2A/H4R3me2s. First, given the abundant occurrence of the motif TCCATT, we referenced the DNA sequence composition of repetitive elements in the human genome. Indeed, TCCATT is part of the consensus sequence of satellite II and III DNA (Table 2.4)^{74,75}, which are types of transcriptionally competent, tandem repetitive elements located primarily in pericentromeric regions⁷⁴.

Satellite Type	Consensus Sequence
satellite II DNA	$[(atTCCATTcg)_2 + (atg)_{1-2}]_n$
satellite III DNA	$[(ATTCC)_{7-13} + (ATTcgggttg)_1]_n$

Table 2.4: Satellite II and III DNA consensus sequences

Subscripts indicate the number of occurrences of a subsequence in the consensus sequence. The motif TCCATT is displayed in uppercase. For satellite III DNA, the motif also appears when two instances of the first subsequence are juxtaposed. Adapted from^{74,75}.

To determine if satellite II and III DNA are the source of the TCCATT motif detected, we analyzed the percentage of the total DNA sequence bound to stable nucleosomes containing various histone modifications that is annotated as satellite II and III DNA (or other repetitive elements; Figure 2.3a). Indeed, around 63% of the total DNA sequence bound to stable H2A/H4R3me2s nucleosomes is satellite II and III DNA, while none of the stable H2A.Z nucleosome-bound DNA is (Figure 2.3a). Satellite II and III DNA also contribute to the DNA sequence bound to stable nucleosomes containing H4K20me3 or H3K9me3, though they comprise a lower percentage (around 7% and 8%, respectively; Fig 2.3a). Thus, stable H2A/H4R3me2s nucleosomal DNA is enriched for TCCATT motifs derived from satellite II and III DNA. As an interesting aside, we find that a substantial portion of the DNA bound to nucleosomes containing stable H4K20me3 or H3K9me3 is retrotransposons; this is not the case for stable nucleosomes containing H2A/H4R3me2s.

Finally, we explored further the relationship between satellite II and III DNA and various histone modifications. For each histone modification, we calculated *occupancy*⁷⁶ over aligned satellite II (or III) DNA sequences, where occupancy is defined as the fraction of sequences at a

position that are bound to a stable nucleosome containing that histone modification (see *Methods*). We found that H2A/H4R3me2s and H4K20me3 had the highest occupancy over satellite II DNA sequences (0.266 and 0.289, respectively) and satellite III DNA sequences (0.159 and 0.142, respectively). H3K9me3 followed closely with occupancies of 0.140 and 0.045 over satellite II and III DNA, respectively. On the other hand, H2A.Z and 18 other histone methylations in the Barski *et al.*⁴⁶ dataset had no or almost no occupancy over these satellites (1 methylation, H3R2me2a, had low occupancy). These findings are depicted in Figure 2.3b.

Thus, downstream analysis functionally relates the elicited features to H2A/H4R3me2s and to each other: H2A/H4R3me2s, H4K20me3, and H3K9me3 all occur on stable nucleosomes in satellite II and III DNA sequences, from which the motif TCCATT is derived. These interactions are consistent with recent biochemical experimental results, a point we return to in the Discussion.

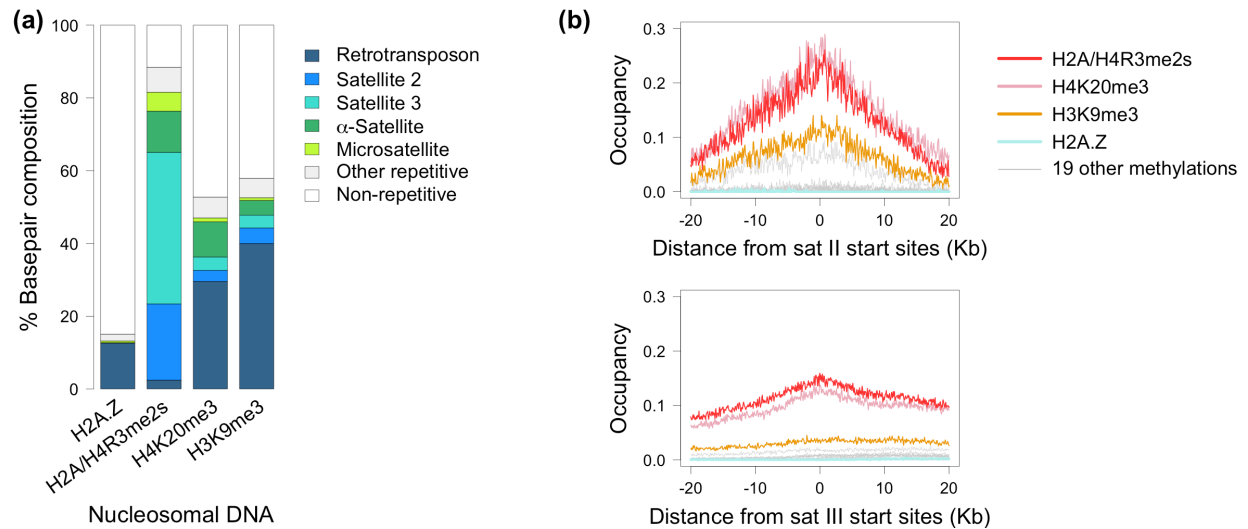


Figure 2.3: Relationship between stable nucleosomes containing histone modifications and satellite II and III DNA sequences

(a) The percentage contributions of types of repetitive elements to the total DNA sequence bound to nucleosomes containing the indicated histone modification. (b) The fraction of start site - aligned satellite II (*upper*) or III (*lower*) DNA sequences occupied by stable nucleosomes containing the indicated histone modification

2.5 Discussion

Emerging, high-throughput epigenomic data, including ChIP-seq data, may provide insight into mechanisms of chromatin structure and gene regulation. However, realizing the full potential of this data requires a computational framework that reduces bias; maximizes algorithm accuracy; and elicits predictive and biologically interpretable features. To this end, we classified nucleosomes containing H2A/H4R3me2s or H2A.Z, as in⁴⁹, but instead employed appropriate data pre-processing and advanced classification algorithms, resulting in greatly improved accuracy and interpretative yield.

Indeed, interpretation of ChIP-seq is challenging because of the magnitude and complexity of the data (issues of quality and pre-processing, aside). This is particularly true when comparing multiple histone modifications (or transcription factors). Encouragingly, approaches aiming to improve ChIP-seq interpretation, albeit not directly applicable to our analyses, appear in the recent literature. For example, Fernandez *et al.*⁷⁷ use a genetic algorithm to identify the optimal number of histone modification profiles to combine to identify transcriptional enhancers, while Beck *et al.*⁷⁸ aim to improve ChIP-seq interpretation by incorporating information about peak shape via linear predictive coding.

In light of these challenges, and given the problems with enumerative feature approaches (e.g., all 6-mers; discussed in detail below), we decided to employ a recently devised pipeline for sequence-based classification, DMFS⁵², that focuses on feature interpretation. DMFS elicits a small set of *a priori* discriminatory features (motifs) using a subsequently withheld data partition. Using DMFS, we evaluated a feature length range between 5 and 10 bp by eliciting < 300 *a priori* discriminatory motifs. In contrast, evaluating this length range with enumerative approaches would require a burdensome, if not prohibitive, $\sum 4^k = 1397760$ features. Thus, feature length often needs to be highly restricted for enumerative approaches, which can then fail to elicit longer, potentially important (interpretable) features. Even with feature length prescriptions, enumerative approaches still employ multitudes of noise features, which can degrade performance⁵³ and complicate determination of feature importance and interpretation. Thus, using DMFS to eliminate univariately unimportant features at the outset has advantages; however, it can miss features whose effects are strict (second or higher order) interactions.

Some attempts have been made to improve interpretation of enumerative feature classification. Most existing enumerative techniques rely heavily on support vector machine

(SVM) classifiers that employ sophisticated, problem-specific kernels, notably the spectrum kernel⁷⁹ and variants thereof^{80,81}, such as the so-called “blended spectrum” kernel used previously⁴⁹ to analyze the data considered here. Determining feature importance for such approaches is arguably very challenging (it is challenging, in general, for SVMs), given inherent feature dependencies (overlaps at neighboring positions) and kernel complexity. Some inventive methods have been developed to address these issues^{82,83}. Nevertheless, these methods are necessarily constrained: input sequences need to be the same length and only select SVM kernels are supported. Thus, another advantage of the DMFS approach is that it provides a modular, all-purpose, pipeline applicable to any (binary) classification problem with any sequence inputs.

In the current study, we employed DMFS for sequence-based classification using pre-processed data. For the sake of comparability, we also tried applying DMFS to *raw*, aligned, extended ChIP-seq reads as used in⁴⁹, which resulted in a classification accuracy similar to that of Gervais and Gaudreau⁴⁹. Thus, while DMFS provided the benefits of ready interpretation, modularity, and computational efficiency, the improvements in performance that we achieved are largely attributable to data pre-processing. Indeed, several authors^{50,51} have advocated ChIP-seq data pre-processing based on observations of bias and extensive background reads. Peak-finding methods have also been specifically designed for histone modification ChIP-seq data: SICER⁸⁴ identifies broad chromatin domains enriched for a histone modification, while NPS⁴⁴ identifies individual, stable nucleosomes that contain a histone modification. Our study is valuable in that it demonstrates empirically the gains in classification performance that result from ChIP-seq data pre-processing, thus substantiating the advocacy thereof.

Another valuable aspect of our study is that the identified features are consistent with recent biochemical experimental results. Our classification approaches identified the motif

TCCATT (derived from satellite II and III DNA sequences) and the histone modifications H4K20me3 and H3K9me3 as predictive of H2A/H4R3me2s nucleosomes. Consistent with this, Zhao *et al.*⁷⁰ recently demonstrated that H4R3me2s provides a direct binding site for the DNA methyltransferase (Dnmt3A) that methylates satellite II and III DNA⁸⁵⁻⁸⁷. The enzyme that mediates H3K9me3 also interacts directly with Dnmt3A⁸⁸. Furthermore, the proper occurrence of H4K20me3 and H3K9me3 has been shown to be partially dependent on Prmt5, the enzyme that mediates H2A/H4R3me2s⁸⁹. Interestingly, the aberrant expression of satellite II and III DNA, which is observed in senescent cells⁹⁰ and epithelial cancers^{87,91}, may promote genomic instability via chromosomal rearrangements⁹². Thus, our finding that H2A/H4R3me2s, H4K20me3, and H3K9me3 occur in stable nucleosomes in satellite II and III DNA sequences *genome-wide* may be consequential in terms of understanding how these genomic elements are normally repressed in healthy, differentiated tissue.

In future work, we will extend our analyses to classifying the 19 other histone modifications in the Barski *et al.*⁴⁶ dataset. This could be realized using an iterative one-against-all approach, which would be more high-throughput (albeit at the potential cost of diluting discriminatory signals), or using a targeted, biologically motivated approach. With respect to the latter, of particular interest would be discriminating between histone modifications that localize with facultative (e.g., H3K27me3) and constitutive (e.g., H3K9me3) heterochromatin. Indeed, DNA elements capable of recruiting the facultative heterochromatin machinery have not been identified in the human genome so far, though they have been in the *Drosophila* genome (i.e., Polycomb Response Elements⁹³). Additionally, we will explore the impact of alternative ChIP-seq normalization approaches, including some more refined, emerging methods⁶⁹. However, because such methods often rely on a mock immunoprecipitation (IP) sample, which many

otherwise rich ChIP-seq datasets lack (including Barski *et al.*⁴⁶), it would be worthwhile to pursue developing a method for identifying the background in datasets with multiple experimental IPs but no mock IP. Similarly, it would be a great advance to develop an algorithm that could identify and remove read buildups that correspond to PCR amplification bias without collapsing “biological” duplicate reads – especially as the latter will be common in newer datasets with very high sequencing depth. Finally, we could pursue, though more ambitious, developing an algorithm for multi-class classification with a similarly discriminatory framework⁵².

2.6 Conclusions

Our study demonstrates that applying advanced classification algorithms to appropriately pre-processed ChIP-seq data results in greatly improved prediction accuracy and feature interpretative yield in genome-wide discrimination between histone modifications. The discriminatory motif feature selection approach that we employed has the added potential to facilitate interpretation of the biological mechanism underlying the classifier performance. Finally, and perhaps most importantly, the findings presented here demonstrate that statistical / machine learning analyses of epigenomic data can identify interpretable, biologically meaningful properties of histone modifications, which have been difficult to study by traditional genetic experimentation.

2.7 Methods

ChIP-seq data pre-processing

The Barski *et al.*⁴⁶ ChIP-seq dataset for 20 histone methylations and H2A.Z in human CD4⁺ T cells was downloaded as BED files of mapped ChIP-seq reads from:

<http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcell.aspx>. In each sample, duplicate reads were collapsed into single reads to eliminate PCR amplification bias^{50,51}. Samples were normalized for unique read number via down-sampling, in order to eliminate bias from sequencing depth variation⁵¹. Stable nucleosomes with statistically significant enrichment over the background were identified, using NPS⁴⁴, for each of the 20 histone methylations and H2A.Z.

NPS extends reads in the 3' direction to 150bp, corresponding to the length of the MNase-digested mononucleosomal DNA^{44,46}. NPS then employs signal sampling and wavelet denoising to improve signal resolution and reduce background, and Laplacian of Gaussian methods to detect peak edges⁴⁴. We only accepted peaks that pass quality control filtering and statistical significance testing, as in⁴⁴, to reduce false positives. Specifically, peaks must have had a width $80\text{bp} \leq w \leq 250\text{bp}$, a strand ratio $s \leq 3$, and a significant number of reads (Poisson $p \leq 1e-05$). For each such nucleosome peak, we extended the midpoint to 147bp for use in classification.

Classification / Feature elicitation

H2A.Z nucleosomes were down-sampled to match the number of H2A/H4R3me2s nucleosomes to create a balanced classification scheme⁷¹. All performance evaluations are based on the mean of ten random samples of H2A.Z nucleosomes to ensure sampling did not impact the results. Classification was performed using random forests⁵⁴, an algorithm that averages over

an ensemble of classification trees. Briefly, each tree is constructed from a bootstrap sample of the data. Unlike conventional trees, where each node is split using the overall most predictive feature, each node in random forest trees is split using the most predictive feature from a subset of features randomly sampled at that node. This additional injection of randomness serves to de-correlate trees in the ensemble, so that subsequent averaging over the ensemble more effectively decreases prediction variance and thereby improves prediction performance⁵⁸. An unbiased estimate of the prediction error rate is obtained as follows: first, for each tree in the ensemble, classify the data points not included in the bootstrap sample for that tree (so-called out-of-bag (OOB) data); then, average the predictions across all trees where a given data point was OOB^{54,58,94}.

Random forests have two primary parameters: for the number of trees, we used $n_{tree} = 500$; and for the subset of features sampled at each node, we used the default classification value $m_{try} = \text{sqrt}(p)$, where p is the number of features. Compared to other classifiers, random forests have the advantage of being relatively resistant to overfitting and relatively insensitive to parameter tuning, as long as n_{tree} is sufficiently large^{54,94}. All reported area under the Receiver Operating Characteristic curve (*auROC*) values are for random forests, though, for thoroughness, classifications were repeated using support vector machines (SVMs); comparable results were obtained. Fitting of both random forests and SVMs made recourse to the corresponding R packages^{94,95} and to the ROCR package⁹⁶.

Classification was performed using two distinct feature types: histone modification features and DNA sequence features. For histone modification features, we used the 19 histone methylations remaining in the Barski *et al.*⁴⁶ dataset, as well as 18 histone acetylations from the Wang *et al.*⁴⁷ dataset, which was generated by the same research group and in the same cell type.

The latter dataset was downloaded from:

<http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellacetylation.aspx>. To create the overlap matrix, an array of length 37 (for 37 histone modification features) was created for each stable H2A/H4R3me2s or H2A.Z nucleosome. Each entry in the array indicates the number of de-duplicated sequence reads for the given feature modification that co-localize with the stable nucleosome boundaries in a strand-specific manner. Specifically, to be scored: ‘+’ strand feature reads must map within ± 50 bp of the 5’ stable nucleosome boundary, and ‘-’ strand feature reads must map within ± 50 bp of the 3’ stable nucleosome boundary.

To generate DNA sequence features, we used DMFS:

<https://bitbucket.org/haoxiong/dmfs-code/>⁵². DMFS elicits a small set of *a priori* informative motifs that discriminate between positive (here, H2A/H4R3me2s) and negative (here, H2A.Z) classes. Unlike enumerative (e.g., all 6-mers) approaches, DMFS avoids the generation of abundant noise features, which can compromise prediction and interpretation⁵³. Additionally, it allows longer, potentially informative features to be evaluated. To avoid data reusage, DMFS requires an additional level of data partitioning, utilizing a *discovery* set for initial discriminatory motif finding and a *classification* set for subsequent random forest (or SVM) analysis. For the fraction of nucleosomal sequences allocated to the discovery set, we used the recommended⁵² value $f = 0.2$; we ultimately evaluated five instances of the data being randomly partitioned as such, to ensure partitioning did not impact the results. A key component of the DMFS pipeline is the tool employed for eliciting discriminatory motifs. We used the default tool – Wordspy^{97,98} – selected in view of its impressive performance in benchmarking studies⁹⁸. Remaining DMFS parameter settings were: minimum motif length $l = 5$, maximum motif length $m = 10$ (with both

DNA strands being searched); and at most $M = 2$ mismatches, when aligning elicited motifs to classification set sequences.

Feature importance and downstream analysis

To identify the most individually predictive features, random forest feature importance was assessed using the mean decrease in Gini index (MDG). Briefly, the Gini index is a measure of statistical impurity. Every time a node is split in a tree, the daughter nodes become more homogenous and, thus, have a lower Gini index than the parent node. A robust measurement of feature importance can be obtained as follows: for each feature, average across all random forest trees the decrease in Gini index that results from splitting a node on that feature⁵⁸. Permutation testing was performed to estimate the statistical significance of variable importance: MDG scores were compared to the distribution of scores from 100,000 classifications using data with permuted class labels.

Downstream analysis was performed for a motif found in many of the elicited sequence features. The genomic coordinates of repetitive DNA sequences were downloaded from the RepeatMasker track of the Table Browser⁹⁹ of the UCSC Genome Browser (build hg18). Based on Repbase Update¹⁰⁰ annotations, satellite II DNA (*repName* = HSATII) and satellite III DNA (*repName* = (CATTC)*n* , (GAATG)*n*) coordinates were extracted. For each histone modification, we calculated the percentage of its total stable nucleosome-bound DNA sequence that consists of satellite II or III DNA. Additionally, for each histone modification, we calculated its *occupancy* along satellite II DNA, or satellite III DNA, sequences aligned by start site — where *occupancy*⁷⁶ is defined as the fraction of sequences bound to a stable nucleosome, in this context, with the histone modification.

2.8 Acknowledgments

Some computations were performed using the UCSF Biostatistics High Performance Computing System. DC was supported in part by NIH Training Grant T32 GM007175. We thank Alain Gervais, Dustin Schones, and Keji Zhao for clarifying details of previous analyses through correspondence. We thank Richard Tabor for assisting with the data storage of Sequence Read Archive files.

Chapter 3: Distance-based assessment of the localization of functional annotations in 3D genome reconstructions

3.1 Citation

Capurso D & Segal MR (2014). Distance-based assessment of the localization of functional annotations in 3D genome reconstructions. *BMC Genomics*, **15**:992.

3.2 Abstract

Background

Recent studies used the contact data or three-dimensional (3D) genome reconstructions from Hi-C (chromosome conformation capture with next-generation sequencing) to assess the co-localization of functional genomic annotations in the nucleus. These analyses dichotomized data point pairs belonging to a functional annotation as “close” or “far” based on some threshold and then tested for enrichment of “close” pairs. We propose an alternative approach that avoids dichotomization of the data and instead directly estimates the significance of distances within the 3D reconstruction.

Results

We applied this approach to 3D genome reconstructions for *Plasmodium falciparum*, the causative agent of malaria, and *Saccharomyces cerevisiae* and compared the results to previous approaches. We found significant 3D co-localization of centromeres, telomeres, virulence genes, and several sets of genes with developmentally regulated expression in *P. falciparum*; and significant 3D co-localization of centromeres and long terminal repeats in *S. cerevisiae*. Additionally, we tested the experimental observation that telomeres form three to seven clusters

in *P. falciparum* and *S. cerevisiae*. Applying affinity propagation clustering to telomere coordinates in the 3D reconstructions yielded six telomere clusters for both organisms.

Conclusions

Distance-based assessment replicated key findings, while avoiding dichotomization of the data (which previously yielded threshold-sensitive results).

3.3 Background

Recent studies^{42,43,45} employed chromosome conformation capture with next-generation sequencing (Hi-C¹⁰¹) to systematically identify genomic regions in physical, three-dimensional (3D) proximity. The resulting contact data lists two genomic positions—each corresponding to a restriction enzyme site—and the frequency with which they were paired-end sequenced together. The smaller the 3D distance between two genomic positions, the larger their interaction frequency should be. Given this relationship, 3D genome reconstructions have been generated from the contact data via constrained optimization for several organisms including *Saccharomyces cerevisiae*⁴³ and the asexual stages of *Plasmodium falciparum*⁴⁵, the causative agent of malaria. Both of these are eukaryotic, haploid, and have relatively small genomes (compared to human). The constraints used in the reconstruction optimization derive from external biological knowledge about genome organization^{43,45}.

Both contact data and attendant 3D genome reconstructions are exciting developments because they provide relatively high resolution, genome-wide information on chromosome organization — which previously could only be probed with low-throughput, low-resolution techniques such as fluorescent in situ hybridization (FISH; contrasted in ¹⁰²). There is now widespread interest in using this data to gain insight into the 3D nuclear localization of

functional genomic annotations (e.g. centromeres, gene ontology (GO) sets). This interest is based on the hypothesis that genome function is linked to its organization⁵⁵. For example, co-regulated genes may be physically co-localized in the nucleus during transcription¹⁰³. Similarly, 3D genome organization likely influences genome stability³⁹ and the location of DNA breakpoints and gene fusions³⁹, including those that drive certain cancers⁴⁰.

Ay et al.⁴⁵ recently assessed the co-localization of functional annotations in *P. falciparum* 3D genome reconstructions; however, their approach led to results that were difficult to interpret. Their assessment was performed as follows. For all data point pairs belonging to a given functional annotation, they dichotomized (Euclidean) distances as “close” or “far” based on prescribed thresholds (10%, 20%, or 40% of the nuclear diameter). Then, they assessed enrichment of “close” pairs in that functional annotation using methods developed for contact data⁵⁵. In the results of this analysis, some functional annotations were significant across all thresholds; however, many functional annotations were significant for only one (or two) threshold(s) but not the other(s). Further, there was often no consistent relationship with respect to threshold. This makes interpretation difficult, especially since it is not obvious what constitutes a good choice for a biologically meaningful threshold. We refer to this approach as “dichotomized distance enrichment” throughout the paper.

Similar analyses have been performed in *S. cerevisiae*^{55,104,105} using contact data rather than the 3D genome reconstruction. Here, pairs of data points belonging to a functional annotation were dichotomized as “close” if they were observed together (i.e. if their interaction frequency passed (False Discovery Rate¹⁰⁶) filtering); otherwise they were “far”. Then, the enrichment of “close” pairs in the functional annotation was tested. We refer to this approach as “dichotomized contact enrichment” throughout the paper.

Rather than dichotomizing the data, we propose directly assessing the significance of distances derived from the 3D reconstruction. This approach is potentially an improvement over previous analyses since it avoids dichotomization of distances (which could incur information loss) and does not require (arbitrary) thresholding or tuning. For a given functional annotation, we computed the median of pairwise Euclidean distances (MPED) between data points belonging to that functional annotation and then assessed the significance of this test statistic by resampling. We also expanded to two-tailed analyses in order to enable tests for *dispersion* of functional annotations since, for example, localization near the nuclear periphery is functionally relevant³⁶. Our approach provided novel findings, replicated key results from prior analyses and provided unambiguous inference for functional annotations that previously reported significance levels that varied by dichotomization threshold. We refer to our approach as “MPED assessment” throughout the paper.

3.4 Results

We performed MPED assessment of functional annotation localization in 3D genome reconstructions (see *Methods*) for *P. falciparum* Ring stage⁴⁵ and *S. cerevisiae*⁴³ from two different restriction enzyme libraries, HindIII and EcoRI (Figure 3.1). We also tested dichotomized contact enrichment (as in⁵⁵; see *Methods*) and compared the results. Results for dichotomized distance enrichment have been reported in detail previously (see Table S5 in⁴⁵).

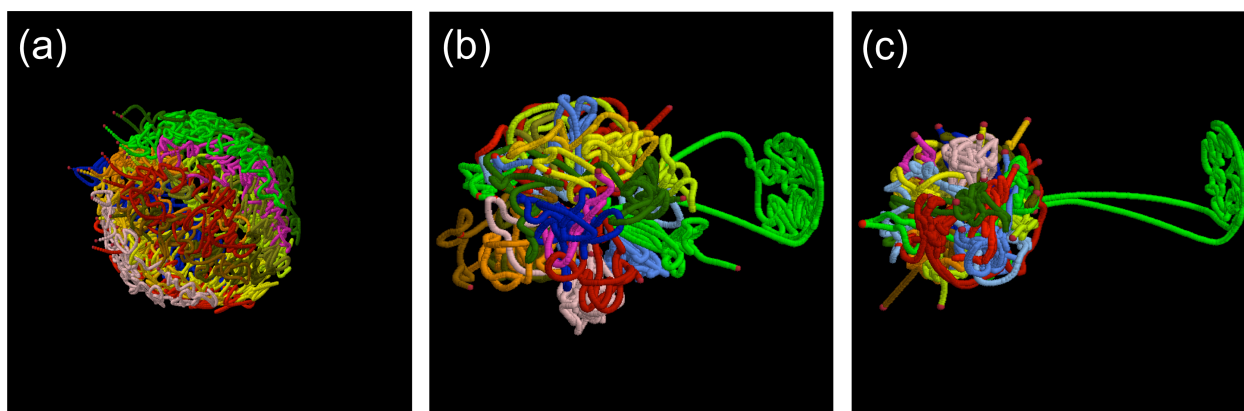


Figure 3.1: 3D genome reconstructions

(a) *P. falciparum* Ring stage 3D genome reconstruction. *S. cerevisiae* 3D genome reconstructions from (b) HindIII or (c) EcoRI restriction enzyme libraries. The diameter of the *P. falciparum* reconstruction is 70 nm. The diameters of the *S. cerevisiae* reconstructions are 200 nm.

3D localization of P. falciparum functional genomic annotations

For *P. falciparum* Ring stage, we assessed the localization of the following functional annotations: centromeres, telomeres, virulence (VRSM) genes, rDNAs, and 15 clusters of genes with developmentally regulated expression^{45,107}. We used normalized¹⁰⁸ *P. falciparum* Ring stage contact data and the (extensively validated) 3D genome reconstruction inferred from these data⁴⁵.

Centromeres, telomeres, and VRSM genes were significantly co-localized under MPED assessment (Table 3.1). These functional annotations were also significantly co-localized under dichotomized contact enrichment (Table 3.1) and under dichotomized distance enrichment at all three thresholds examined (10%, 20%, or 40% of the nuclear diameter; see Table S5 in⁴⁵).

Furthermore, experimental FISH data supports the nuclear clustering of telomeres in *P.*

falciparum^{109,110}.

Eight out of 15 clusters of genes with developmentally regulated expression (including several with Ring stage expression) were significantly co-localized under MPED assessment, but only 1 was significantly co-localized under dichotomized contact enrichment (Table 3.1). Of the 8 expression clusters significantly co-localized under MPED assessment, only 2 were significant across all three thresholds under dichotomized distance enrichment (see Table S5 in ⁴⁵); the other 6 had threshold-dependent significance under dichotomized distance enrichment (and thus ambiguous interpretation in the previous study ⁴⁵). In the *Discussion*, we comment on why assessing localization at the 3D reconstruction level (with MPED) may reveal significant co-localization for some functional groups that was not detected using contact level data .

Functional annotation	MPED q-values	Contact enrichment q-values
Centromeres	6.0e-05	8.4e-05
Telomeres	6.0e-05	8.4e-05
VRSM (all)	6.0e-05	8.4e-05
VRSM (subtelomeric)	6.0e-05	8.4e-05
VRSM (internal)	1.6e-04	8.4e-05
rDNA genes	0.42	0.10
Cluster 1	0.73 ↓	0.17
Cluster 2	4.4e-02	0.70 ↓
Cluster 3	0.18	0.45 ↓
Cluster 4 (Ring)	6.0e-05	1.0e-02
Cluster 5 (Ring)	0.24	0.45 ↓
Cluster 6 (Ring)	6.0e-05	0.70 ↓
Cluster 7 (Ring)	6.0e-05	0.39 ↓
Cluster 8	4.0e-02	0.86 ↓
Cluster 9	1.0e-02	0.39 ↓
Cluster 10	2.1e-03	0.81
Cluster 11	0.10	0.74 ↓
Cluster 12	9.2e-03	0.11 ↓
Cluster 13	6.5e-02	0.44 ↓
Cluster 14	0.11	0.70 ↓
Cluster 15	5.2e-02	0.81 ↓

Table 3.1: Assessment of the 3D localization of functional annotations in *P. falciparum* Ring stage

MPED: the median of pairwise Euclidean distances in the 3D reconstruction. *Contact enrichment*: enrichment of dichotomized “close” pairs in the Hi-C contact data. Gray shading indicates q-value <0.05. Down arrow indicates dispersion (otherwise co-localization). All functional annotations that were tested are included. “Cluster N” refers to genes with life cycle - regulated expression, which were clustered in Le Roch et al.¹⁰⁷. Clusters that have high gene expression in the Ring stage are indicated in parentheses.

3D localization of S. cerevisiae functional genomic annotations

For *S. cerevisiae*, we assessed the localization of 264 GO terms and 17 other functional annotations, including centromeres, telomeres, retrotransposon long terminal repeats (LTRs), classes of non-coding RNAs, classes of replication origins, classes of DNA breakpoints, and classes of cell cycle -regulated genes (full list in *Methods*). We report functional annotations that were significant under MPED assessment with both restriction enzyme libraries (HindIII and EcoRI) or significant with both libraries under dichotomized contact enrichment.

There is no indication that the *S. cerevisiae* Hi-C data was normalized in previous studies^{43,55} prior to generating the 3D genome reconstructions or assessing functional annotation localization: the original study⁴³ preceded the formalization of Hi-C data normalization pipelines^{108,111,112} that redress biases due to factors such as fragment length, GC content and mappability. Accordingly, we normalized the *S. cerevisiae* Hi-C contact data (see *Methods*) and then generated new reconstructions, as in⁴³, from the normalized contact data (Figure 3.1) before assessing functional annotation localization.

Centromeres and LTRs were significantly co-localized under MPED assessment and under dichotomized contact enrichment (Table 3.2). Previous analyses of this *S. cerevisiae* Hi-C data also found significant co-localization of centromeres⁵⁵ and LTRs¹¹³. Furthermore, experimental FISH data support the nuclear clustering of centromeres¹¹⁴ and LTRs¹¹⁵ in *S. cerevisiae*. Several GO terms that map to LTRs (e.g., retrotransposon nucleocapsid, transposition) were also significantly co-localized under both analyses but are not included in Table 3.2 because of the redundancy in the mapping.

Telomeres were significantly co-localized under dichotomized contact enrichment, but not under MPED assessment (Table 3.2). Experimental FISH data support nuclear clustering of

S. cerevisiae telomeres^{116,117}. In the *Discussion*, we comment on why assessing localization at the 3D reconstruction level (with MPED) may not detect significant co-localization for some functional groups that were detected at the contact data level (particularly the difficulty of generating a null distribution for telomeres).

The previous study that analyzed *S. cerevisiae* functional annotation localization under dichotomized contact enrichment reported significant co-localization of certain functional groups (e.g., early replication origins (Clb5 and Rad53), and tRNAs)⁵⁵ that were not replicated in our analysis under dichotomized contact enrichment. This difference may be due to our testing a much larger number of functional groups (and the corresponding multiplicity correction) and/or our normalization of the data prior to assessment. Experimental FISH data supports tRNA clustering in *S. cerevisiae*¹¹⁸. Under dichotomized contact enrichment, our q-values for tRNAs were 2.4e-02 (HindIII) and 0.55 (EcoRI). Under MPED assessment, our q-values for tRNAs were 0.64 (HindIII) and 2.0e-03 (EcoRI).

Functional annotation	MPED q-values		Contact enrichment q-values	
	HindIII	EcoRI	HindIII	EcoRI
Centromeres	4.0e-04	3.7e-04	2.8e-03	5.6e-03
Long terminal repeats	4.0e-04	3.7e-04	2.8e-03	1.9e-02
Telomeres	0.86 ↓	0.13 ↓	5.0e-02	5.6e-03

Table 3.2: Assessment of the 3D localization of functional annotations in *S. cerevisiae*

MPED: the median of pairwise Euclidean distances in the 3D reconstruction. *Contact enrichment*: enrichment of dichotomized “close” pairs in the Hi-C contact data. Gray shading indicates q-value <0.05. Down arrow indicates dispersion (otherwise co-localization). Functional annotations are included if they were significant for both restriction enzyme libraries (HindIII and EcoRI) in either analysis.

Generating a null referent distribution

In our MPED assessment of functional annotation localization above, we generated a null referent distribution by resampling points from the same chromosome as observed (i.e. preserving the chromosome structure of the data). An alternative approach is to resample preserving the distance that a data point is from the center of the nucleus (within a range), but not preserving the chromosome structure. Such a resampling scheme may detect functional groups that are co-localized given the Rabl configuration of the *S. cerevisiae* 3D genome reconstructions⁴³. To perform such a resampling scheme, we divided the radius of the nucleus into fifths and created a series of concentric spheres at each partition. Points were then resampled from the 3D annulus (ring) between concentric spheres. The results under MPED assessment with annulus resampling were similar to those with chromosome resampling for both organisms (Tables 3.3 and 3.4).

Functional annotation	Chromosome resampling q-values	Annulus (Rab1) resampling q-values
Centromeres	6.0e-05	7.0e-05
Telomeres	6.0e-05	7.0e-05
VRSM (all)	6.0e-05	7.0e-05
VRSM (subtelomeric)	6.0e-05	7.0e-05
VRSM (internal)	1.6e-04	7.0e-05
rDNA genes	0.42	0.22
Cluster 1	0.73 ↓	0.77 ↓
Cluster 2	4.4e-02	2.6e-03
Cluster 3	0.18	3.3e-04
Cluster 4 (Ring)	6.0e-05	2.6e-04
Cluster 5 (Ring)	0.24	3.5e-03
Cluster 6 (Ring)	6.0e-05	7.0e-05
Cluster 7 (Ring)	6.0e-05	7.9e-02
Cluster 8	4.0e-02	0.45
Cluster 9	1.0e-02	2.7e-03
Cluster 10	2.1e-03	6.3e-03
Cluster 11	0.10	2.5e-02
Cluster 12	9.2e-03	1.2e-04
Cluster 13	6.5e-02	1.0e-03
Cluster 14	0.11	3.8e-04
Cluster 15	5.2e-02	0.22

Table 3.3: Comparison of resampling schemes for distance-based assessment of the localization of functional annotations in *P. falciparum* Ring stage

Points were resampled within the same chromosome or within the same annulus. Gray shading indicates q-value <0.05. Down arrow indicates dispersion (otherwise co-localization). All functional annotations that were tested are included. “Cluster N” refers to genes with life cycle - regulated expression, which were clustered in Le Roch et al.¹⁰⁷. Clusters that have high gene expression in the Ring stage are indicated in parentheses.

Functional annotation	Chromosome resampling q-values		Annulus (Rabl) resampling q-values	
	HindIII	EcoRI	HindIII	EcoRI
Centromeres	4.0e-04	3.7e-04	5.1e-04	0.21
Long terminal repeats	4.0e-04	3.7e-04	5.1e-04	4.0e-04
Telomeres	0.86 ↓	0.13 ↓	0.88 ↓	0.39 ↓

Table 3.4: Comparison of resampling schemes for distance-based assessment of the localization of functional annotations in *S. cerevisiae*

Points were resampled within the same chromosome or within the same annulus. Gray shading indicates q-value <0.05. Down arrow indicates dispersion (otherwise co-localization). Functional annotations from Table 2.2 are shown.

Affinity propagation clustering applied to 3D telomere coordinates

Experimental FISH data indicate that telomeres form 4 to 7 clusters in *P. falciparum*^{109,110} and 3 to 7 clusters in *S. cerevisiae*^{116,117}. To determine if we could recapitulate this property of telomere organization from the 3D genome reconstructions (and to identify which telomeres are close to each other) we applied affinity propagation (AP) clustering¹¹⁹ to telomere coordinates in the 3D genome reconstructions. Unlike many other clustering algorithms (e.g. *k*-means) where the number of clusters needs to be specified from the outset, AP clustering optimizes the number of clusters within the algorithm. Applying AP clustering yielded 6 telomere clusters for both *P. falciparum* (Figure 3.2) and *S. cerevisiae* (Figure 3.3), consistent with the FISH data. This also revealed which telomeres are close to each other in the 3D genome reconstructions (Figures 3.2 and 3.3).

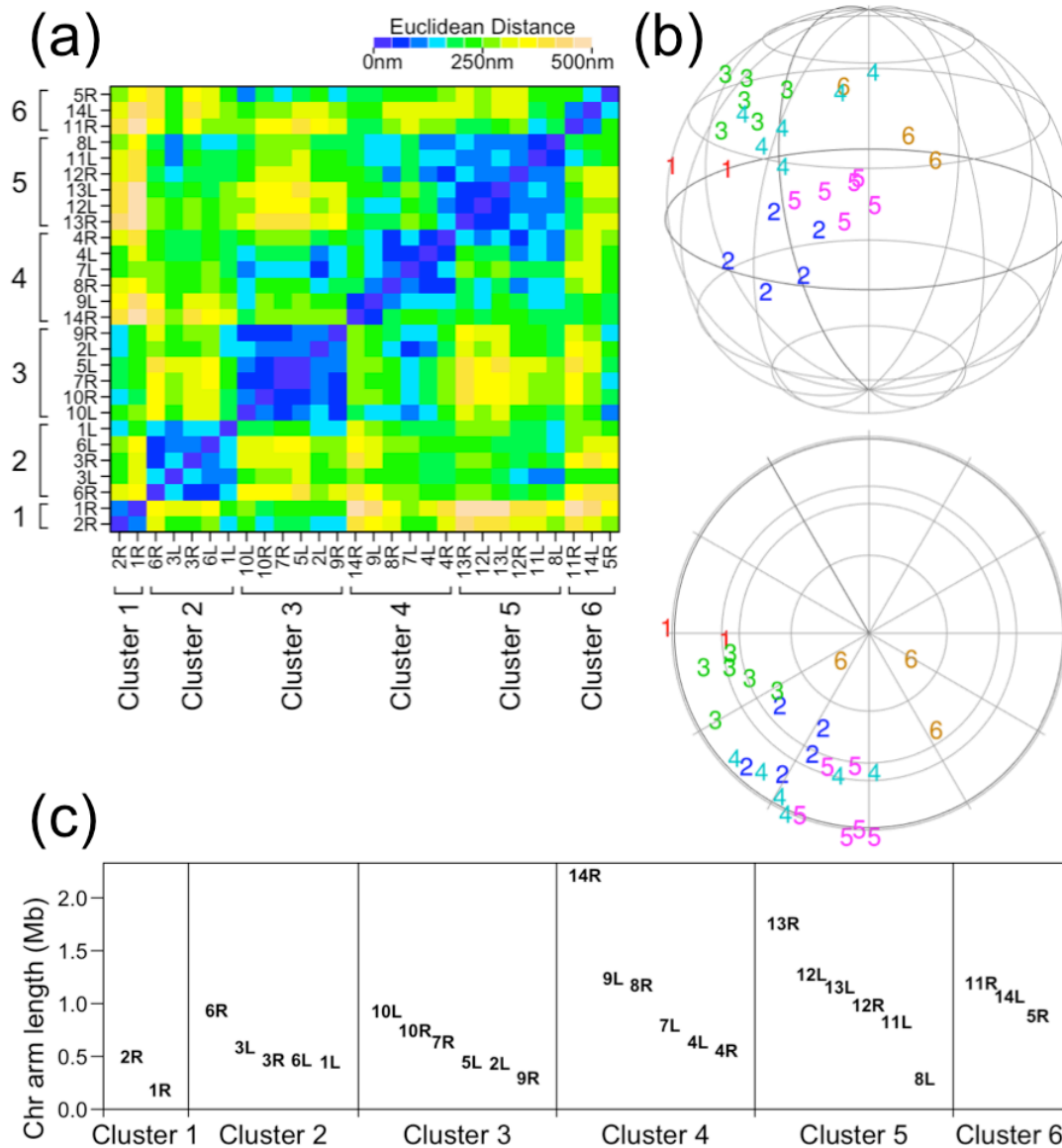


Figure 3.2: Affinity Propagation clustering applied to 3D telomere coordinates for *P. falciparum* Ring stage

(a) Heat map of Euclidean distances between telomeres. The clustering is indicated. (b) Positions of telomeres in the 3D reconstruction plotted as the cluster number. *Upper*: side view. *Lower*: top view, a 90-degree rotation forward about the z-axis relative to the side view. (c) The chromosome arm lengths of telomeres in each cluster.

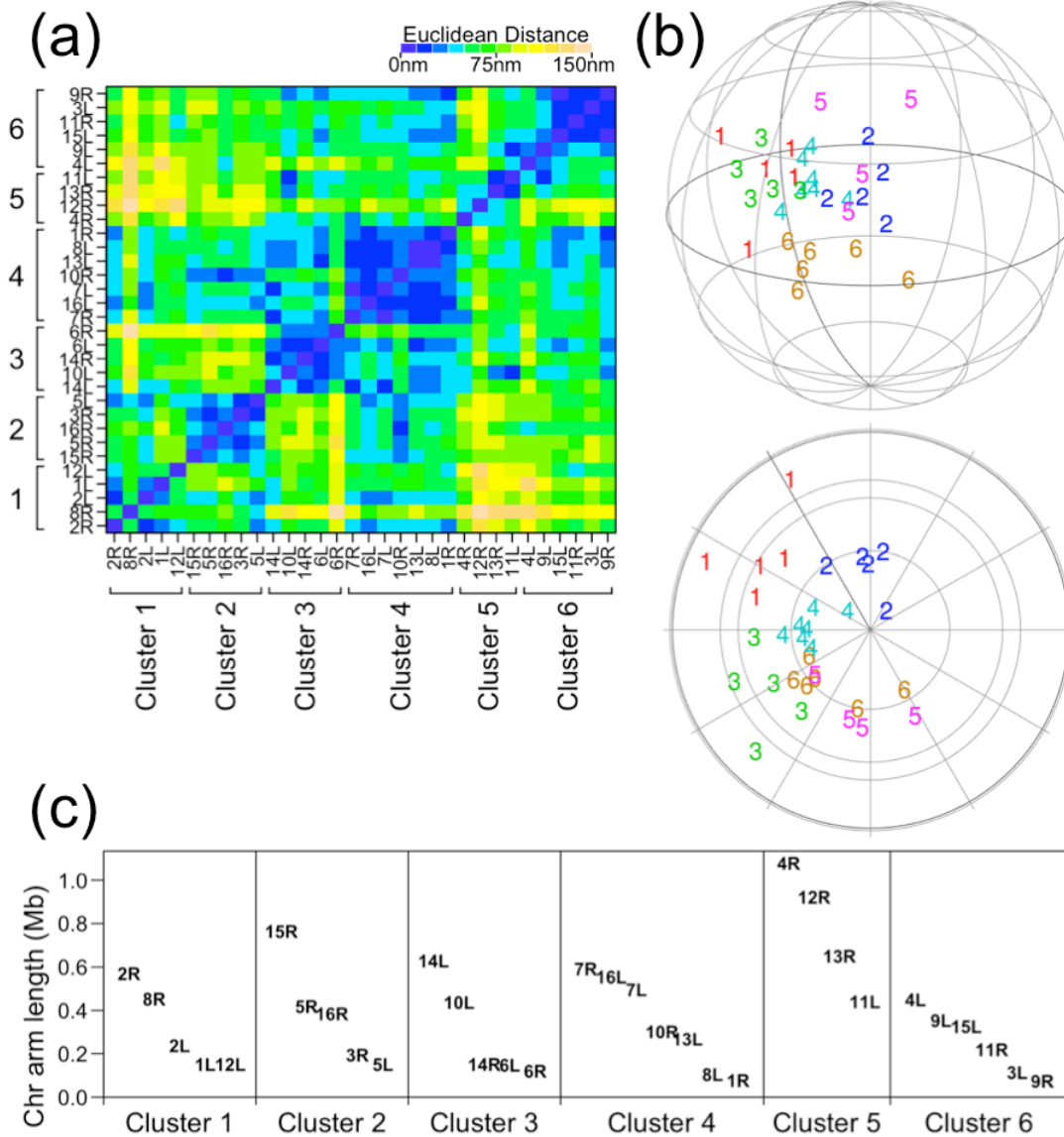


Figure 3.3: Affinity Propagation clustering applied to 3D telomere coordinates for *S. cerevisiae* (HindIII)

(a) Heat map of Euclidean distances between telomeres. The clustering is indicated. (b) Positions of telomeres in the 3D reconstruction plotted as the cluster number. *Upper*: side view. *Lower*: top view, a 90-degree rotation forward about the z-axis relative to the side view. (c) The chromosome arm lengths of telomeres in each cluster.

3.5 Discussion

In this study, we propose using MPED to assess functional annotation localization and applied this approach to *P. falciparum* and *S. cerevisiae* 3D genome reconstructions. We argue that, if functional annotation localization assessment is to be performed at the 3D genome reconstruction level, then MPED assessment offers advantages over dichotomized distance enrichment⁴⁵ because it avoids dichotomization of the data (which could incur information loss) and does not require (arbitrary) thresholding or tuning thereby providing unambiguous results.

However, as with any statistic and associated inferential assessment, MPED embodies specific choices and assumptions. For the statistic, we have employed the *median* (because of its robustness and resistance properties) of *all* pairwise distances (because this does not require tuning as, for example, would be necessary with k nearest neighbor distances). Evaluation of alternative formulations (mean rather than median; k nearest neighbor distances rather than pairwise distances) had comparable results (when $k \geq 2$). For inferential assessment, we have used two approaches to generating null referent distributions (as described above); other potentially organism-specific possibilities could be entertained. A strength of methods imposing dichotomization is that obtaining reasonable referent distributions is relatively straightforward.

There are other putative advantages of assessing functional annotation localization at the 3D reconstruction level: (i) while the contact data is inherently pairwise, the 3D reconstructions exploit higher order relationships; (ii) the 3D location of sites for which there is missing contact data is readily determined from neighbouring points in the reconstruction because of chromatin contiguity; and (iii) biological and biophysical constraints about genome organization are imposed (e.g. avoidance of steric clashes). Thus, emergent properties of the 3D reconstructions

may reveal significant co-localization of some functional annotations that were not co-localized in the (pairwise) contact data (e.g. *P. falciparum* gene expression clusters).

The advantage of assessing functional annotation localization at the contact data level is that resampling to generate a null distribution makes recourse to only chromosome labels, while at the 3D reconstruction level, resampling makes recourse to the (more complex) chromatin structure. The 3D reconstructions for *S. cerevisiae* have low chromatin density near the nuclear periphery and large chromatin voids in the nucleus (Figure 3.1). Given that *S. cerevisiae* telomeres are in the periphery, resampling making recourse to the chromatin structure thus samples points from more internally than the telomeres de facto (even with annulus resampling), which may make it difficult to detect co-localization. Resampling points *without* making recourse to the chromatin structure (i.e. any X,Y,Z coordinate within an annulus) would not be stringent enough. *S. cerevisiae* telomeres may be co-localized given a spherical 3D nucleus (and ignoring the chromatin structure within); however, MPED assessment does not detect significant co-localization of *S. cerevisiae* telomeres by generating a null distribution by resampling points making recourse to the (complex) chromatin structure.

It is important to note that there are caveats to the use of Hi-C data (whether at the contact data level or 3D genome reconstruction level). Most current Hi-C data represents averages over many cells. The first example of single cell Hi-C in mouse has recently been reported¹²⁰; however, a 3D mammalian genome reconstruction has not yet been generated for computational reasons. Mammalian Hi-C analysis is complicated further by diploid genomes, though methods related to Hi-C have been developed for deconvolving sequence data for homologous chromosomes¹²¹. Finally, Hi-C is a snapshot of highly dynamic chromatin organization; these dynamics are important to understand, but difficult to capture. For the 3D

reconstruction-based approach to be meaningful requires that the reconstruction provides an adequate representation of dynamics and between-cell variation. Methods for making such assessments and devising and contrasting reconstruction algorithms are active research areas¹²²⁻

124

In the current study, we assessed the 3D localization of genomic annotations (point data). Each data point has an X,Y,Z coordinate; co-localization is assessed by estimating the significance of distances between points. In future research, we will expand to assessing the 3D localization of continuous, functional genomic data – for example, by overlying chromatin immunoprecipitation sequencing (ChIP-seq) peak height on top of the 3D reconstructions. While our current research provides a framework for such an analysis, future research will require developing and/or applying methodology suited to detect co-localization of data that has an X,Y,Z coordinate paired with a continuous outcome (peak height).

3.6 Conclusions

When assessing functional annotation localization at the 3D reconstruction level: MPED assessment, as proposed and applied here, offers advantages over the existing approach (dichotomized distance enrichment). MPED assessment replicated key findings from previous analyses, as well as providing novel results, and provided unambiguous significance estimates for functional annotations that previously had significance levels that varied by threshold.

3.7 Methods

***P. falciparum* data and annotations**

The *P. falciparum* Ring stage contact data and 3D reconstruction were obtained at: <http://noble.gs.washington.edu/proj/plasmo3d/>. This data has already been normalized and filtered⁴⁵. Various functional annotations were assessed: centromeres, telomeres, rDNA genes, VRSM genes, and developmentally regulated gene expression clusters¹⁰⁷. All of these annotations are available at the same link as for the *P. falciparum* contact data (above).

***S. cerevisiae* data and annotations**

S. cerevisiae contact data (pre-FDR, no masking) for HindIII and EcoRI⁴³ were obtained at: <http://noble.gs.washington.edu/proj/yeast-architecture/sup.html>. We normalized this contact data for GC content, mappability, and fragment length by applying HiCNorm¹¹² genome-wide (chromosome by chromosome). We then filtered to retain the top contacts by interaction frequency. We generated new 3D genome reconstructions⁴³ for HindIII and EcoRI based on this normalized and filtered contact data.

Various functional annotations were assessed. Annotations for centromeres, telomeres, retrotransposon long terminal repeats (LTRs), transfer RNAs (tRNAs) and small nucleolar RNAs (snoRNAs) were obtained from the Table Browser of the UCSC Genome Browser⁹⁹.

Annotations for early Clb5-independent replication origins, late Clb5-dependent replication origins, early Rad53-regulated origins, and late Rad53-regulated origins from¹²⁵ were obtained at: <http://noble.gs.washington.edu/proj/yeast-architecture/sup.html>. Gene Ontology (GO) term annotations were obtained from the Gene Ontology Website¹²⁶ and corresponding gene coordinates were obtained from the Table Browser of the UCSC Genome Browser⁹⁹. We filtered

GO terms by membership: 264 terms with between 25 and 120 genes were retained for analysis.

Cell cycle-regulated genes (5 clusters of genes with expression that peaks during M/G1, G1, S, S/G2, or G2/M) from ¹²⁷ were obtained at:

<<http://genome-www.stanford.edu/cellcycle/data/rawdata/>>. Annotations for DNA breakpoints from ¹²⁵ were obtained at: <<http://gbe.oxfordjournals.org/content/1/350/suppl/DC1>>. Genomic positions in these files were for the sc1 assembly of the *S. cerevisiae* genome, so we converted to sc2 assembly positions using the Batch Coordinate Conversion (liftover) tool from the UCSC Genome Browser¹²⁸. Three categories of DNA breakpoints were used in the analyses: experimentally-induced (mutagenized) breakpoints, evolutionary breakpoints compared to *Kluyveromyces waltii*, and evolutionary breakpoints compared to the hypothetical/inferred ancestor that *S. cerevisiae* and *K. waltii* share^{125,129}.

MPED assessment

The 3D genome reconstruction data consists of a series of “beads” spaced throughout the linear genome. Each bead has a genomic position and a 3D coordinate (X,Y,Z). To map functional annotations to the 3D reconstruction data, we assigned each centromere, for example, to its nearest bead in linear, genomic space.

We assessed functional annotation localization at the 3D genome reconstruction level as follows. We employed the median of pairwise Euclidean distances (MPED) –applied interchromosomally, in order to avoid detection of annotations simply clustered in linear, genomic space⁵⁵. To estimate MPED significance, we generated a null referent distribution by resampling 1e05 times with preservation of the chromosome structure of the data. For example,

for centromeres—where there is one centromere per chromosome—we randomly selected one bead from each chromosome during each resampling, and computed and saved the MPED.

Results from preservation of the chromosome *arm* structure of the data (not shown) were very similar to those obtained from preserving the chromosome structure of the data. We also tried preserving the annulus structure of the data – in other words, preserving the approximate distance that a bead is from the center of the nucleus, but not preserving the chromosome structure of the data. For annulus resampling, we divided the radius into fifths and created concentric spheres at each partition; we then resampled beads from the appropriate annulus (ring) between concentric spheres.

We estimated p-values as follows. When the test statistic was greater than the mean of the null referent distribution (of MPEDs from resampling), the p-value was based on comparison to the upper tail of the distribution (and, if significant, would indicate dispersion). When the statistic was less than the mean of the null referent distribution, the p-value was based on comparison to the lower tail of the distribution (and, if significant, would indicate co-localization). We used False Discovery Rate (FDR)¹⁰⁶ for multiple testing corrections and accepted an FDR q-value of <0.05 as significant.

Dichotomized contact enrichment

The contact data lists two genomic positions— each corresponding to restriction enzyme site (or bin, if the data is binned) — and the frequency with which the two interact (are sequenced together). The normalized contact data was filtered to retain only the top contacts by interaction frequency⁴³. We mapped functional annotations to the filtered contact data as in⁴³: for a given centromere, for example, all restriction sites within a window are assigned to that

centromere (along with the attendant contact data). The window sizes were 5kb for *S. cerevisiae* and 10kb for *P. falciparum*, in line with the resolution/binning of the respective 3D reconstructions^{43,45}.

To assess functional annotation localization from the contact data, we used dichotomized contact enrichment⁵⁵. Pairs of elements belonging to a functional annotation were considered “close” if the restriction enzyme sites to which they map were present together in the filtered contact data. The test statistic is the (genome-wide) ratio of the number of observed, interchromosomal “close” pairs (k) to the number of possible, interchromosomal pairs (m). To estimate $k:m$ significance, we generated a null referent distribution by resampling 1e05 times as follows. For each chromosome, we resampled the same number of restrictions sites as were assigned on that chromosome and then computed and saved the statistic. We estimated p-values by comparing the test statistic to the null referent distribution, as described above for the reconstruction-based assessment. Our analysis differs from⁵⁵ in that we perform a two-tailed assessment. We again used FDR for multiple testing correction with a q-value of <0.05 accepted as significant.

3.8 Acknowledgments

Some computations were performed using the UCSF Biostatistics High Performance Computing System. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1144247 and National Institutes of Health Grant R01 GM109457. DC was supported in part by the National Institutes of Health Training Grant T32 GM007175.

Chapter 4: Identifying hotspots in functional genomic data superposed on 3D chromatin configuration reconstructions

4.1 Citation

Capurso D, Bengtsson H, Segal MR. Identifying hotspots in functional genomic data superposed on 3D chromatin configuration reconstructions. In review.

4.2 Abstract

Background

The spatial organization of the genome influences cellular function, notably gene regulation. Recent studies have assessed the three-dimensional (3D) co-localization of functional annotations (e.g. long terminal repeats) using 3D genome reconstructions from Hi-C (genome-wide chromosome conformation capture) data; however, corresponding assessments for (continuous) functional genomic data are lacking. Here we advance such techniques. We overlaid a *Saccharomyces cerevisiae* 3D genome reconstruction with three chromatin immunoprecipitation-sequencing (ChIP-seq) inputs and contrasted two algorithms for identifying regions in 3-space — “3D hotspots” – for which mean ChIP-seq peak height is significantly elevated: k -Nearest Neighbor (k -NN) regression and the Patient Rule Induction Method (PRIM).

Results

The ChIP-seq inputs were: Swi6, a transcription factor; RNA polymerase II phosphorylated at serine 5 (Pol2Ser5p), the active transcriptional machinery; and Tup1, a repressor. For each input, the algorithms identified a significant and corresponding 3D hotspot. For Swi6, the hotspot contains MSB2 and ERG11 – known Swi6 target genes on different

chromosomes. For Pol2Ser5p, the hotspot contains regions from multiple chromosomes with genes that function in vacuole transport and RNA binding/processing. For Tup1, the hotspot contains NRG1 and YAP6 – known Tup1-regulated genes.

Conclusions

Both algorithms identified significant, corresponding and biologically meaningful 3D hotspots containing distal genomic regions. k -NN regression has the advantage of being invariant to rotations of the (coordinate-free) 3D reconstruction, while PRIM is arguably more robust to parameter tuning and sensitivity to orientation can be assessed over (disparate) rotations. 3D reconstructions are critical to such hotspot detection: attempting to find them using precursor Hi-C contact data is computationally prohibitive.

4.3 Background

The three-dimensional (3D) configuration of chromosomes within the eukaryotic nucleus is consequential for several cellular functions including gene expression regulation and epigenetic patterning¹³⁰ and is also strongly associated with translocation events and cancer driving gene fusions^{16,40}. While visualization of such architecture remains limited to low-resolution, low-throughput, targeted techniques such as Fluorescent In Situ Hybridization (FISH)¹⁰², the ability to *infer* structures at high resolution has been enabled by recently-devised assays derived from chromosome conformation capture (3C) techniques¹³¹. In particular, when coupled with next generation sequencing, such methods (hereafter termed *Hi-C*^{42,43}) yield an inventory of genome-wide chromatin interactions which, in turn, provide a basis for reconstructing 3D configurations, as described below. Such 3D reconstructions are crucial for

discovering functional nuclear compartments^{35,36,132,133}, since without such a guiding structure the search space is prohibitively large.

The contact data from Hi-C analysis lists two genomic positions — each corresponding to a restriction enzyme site (or bin if the data are binned) – and an “interaction frequency”: the number of times the two positions were ligated and paired-end sequenced together. This interaction frequency is inversely related to the physical 3D distance between the two genomic positions in the nucleus^{43,45}.

By quantifying the relationship between interaction frequency and physical distance, Duan et al.⁴³ proceeded to generate a 3D reconstruction of the *Saccharomyces cerevisiae* genome (16 chromosomes, 12.2 Megabases (Mb), and ~6,275 genes) by solving a multi-dimensional scaling criterion^{43,45,122,123,134} via constrained optimization – with constraints based on prior biophysical and biological knowledge (e.g. imposition of within-chromosome contiguity, and avoidance of steric clash). A 3D genome reconstruction has also been generated⁴⁵ for *Plasmodium falciparum* 3D7 (14 chromosomes, 23.3 Mb, and ~5,300 genes), the causative agent of malaria, using a similar approach. Additional methods for generating 3D genome reconstructions using alternate approaches to inferring distances from interaction frequencies and differing optimization methods have been advanced^{122,123,134}, as have methods for gauging the concordance of 3D genome reconstructions¹²⁴.

Several recent studies have used the contact data⁵⁵, the 3D genome reconstructions⁴⁵, or both¹³⁵ to test the hypothesis that functionally related genomic annotations co-localize in 3-space the nucleus. Centromeres, telomeres, and long terminal repeats were detected as significantly co-localized in *S. cerevisiae*¹³⁵. Interestingly, in *P. falciparum*, sets of genes with developmentally regulated expression were detected as significantly co-localized in 3D-reconstruction-based

assessments but not in contact-based assessments¹³⁵. This finding illustrates a potential advantage of 3D reconstructions: they enable the detection of multi-level co-localizations (i.e. of multiple (inter)chromosomal regions), whereas the contact data are inherently limited to detecting strictly pairwise co-localization. For example, reconstruction-based analyses may detect if a set of elements occupies a smaller subset of the nucleus than expected by chance even if no two individual elements are exceptionally close together. Another advantage of the 3D reconstruction is that for genomic regions that have missing contact data, their position in the 3D genome reconstruction is inferred from neighboring genomic regions via chromatin contiguity.

Here, we extend such downstream, functional analyses of 3D genome reconstructions to high-throughput, functional genomic data. We started by (separately) overlaying an *S. cerevisiae* 3D genome reconstruction with the peak height of three chromatin immunoprecipitation-sequencing (ChIP-seq) inputs from¹³⁶: Swi6, RNA polymerase II phosphorylated at serine 5 (Pol2Ser5p), and Tup1.

One previous study superposed high-throughput functional genomic data (microarray gene expression data) on a model-based 3D structure⁵⁹; however, this was solely for visualization purposes. Another used an unsupervised learning approach to identify gene expression profiles that exhibit (global) coherence with the 3D reconstruction⁴⁵. Our study makes the novel contribution of adapting, applying and comparing select supervised learning techniques for analyzing 3D genome reconstructions overlaid with high-throughput functional genomic data with the objective of eliciting focal regions in 3-space – “3D hotspots” – for which the overlaid outcome is extreme. An important motivation for searching for focal 3D hotspots is that downstream analyses of their gene membership can then reveal valuable biological information in contrast to global assessments.

The outcomes analyzed here derive from ChIP-seq as mentioned; however the methods can be applied irrespective of outcome type. The two methods used for identifying 3D hotspots are k -Nearest Neighbor (k -NN) regression^{56,58} and the Patient Rule Induction Method (PRIM)^{58,137}. As noted in the Conclusions, few if any existing methods seem suited to this task.

4.4 Results and Discussion

Data normalization and integration

We normalized the *S. cerevisiae* contact data from⁴³ using HiCNorm¹¹² (see “Methods”) and then generated a new 3D genome reconstruction from the normalized contact data via the constrained optimization approach from⁴³ as per¹³⁵ (the original study⁴³ preceded the formalization of pipelines for normalizing Hi-C contact data^{108,111,112,138}).

Next, we aligned to the reference genome the sequencing reads for the ChIP-seq inputs (Swi6, Pol2Ser5p, and Tup1) and the mock IP control, and then \log_2 -normalized the signal of each ChIP-seq input to the control (see “Methods”). We applied quality control filters, performed read de-duplication, and obtained residuals from smoothing the signal along each chromosome arm (see “Methods”) – this constitutes the final “ChIP-seq peak height” that we proceeded to analyze.

We (separately) superposed the ChIP-seq peak height of each input on the 3D genome reconstruction (Figure 4.1). The 3D genome reconstruction consists of a series of “beads” spaced along each chromosome, with each bead having a genomic position and an (X,Y,Z) coordinate. For each ChIP-seq input, we binned the peak height data at the same genomic spacing (resolution) as that of the beads (see “Methods”). The result is that each bead now has a genomic position, an (X,Y,Z) coordinate, and a ChIP-seq peak height value. We applied k -NN regression

and PRIM to these data, with ChIP-seq peak height as outcome and (X,Y,Z) coordinates as covariates, to identify 3D hotspots.

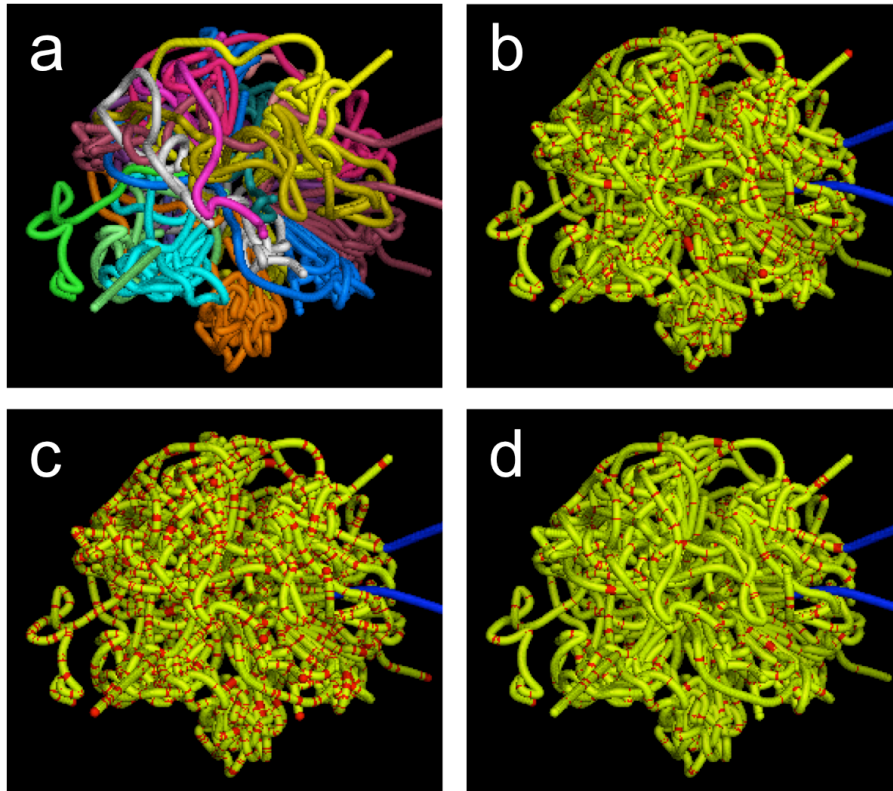


Figure 4.1: ChIP-seq peak height superposed on the 3D chromatin configuration reconstruction

The 3D reconstruction is colored by (a) chromosome, or by ChIP-seq peak height for (b) Swi6, (c) Pol2Ser5p, and (d) Tup1. For (b-d), regions are colored *red* if their \log_2 -normalized ChIP-seq peak height is greater than 1, otherwise they are colored *yellow* (except for the masked chr12 rDNA repeat region, which is colored *blue*).

k-NN regression and PRIM

We adapted k -NN regression to the task of identifying 3D hotspots as follows. At fixed intervals along each chromosome (*seed_interval*), a bead is selected and is grouped with the k beads closest to it in Euclidean distance (see “Methods”). The mean ChIP-seq peak height is computed for each group and the groups are then ranked by mean ChIP-seq peak height. We performed inference as follows. We saved the top 10 hotspots by mean ChIP-seq peak height. Then, we permuted the ChIP-seq peak height values along each chromosome and re-ran k -NN regression on the permuted data and again saved the top 10 hotspots. We repeated the permutation process many times, and then compared the mean ChIP-seq peak height of the top 10 hotspots for the observed data to that of the top 10 hotspots for the permuted data – comparing across each rank (see “Methods”).

The key tuning parameter is k , which governs the extent of local averaging, which we tuned as follows: take the setting (out of $k = 25, 50, 75, 100, 125, \text{ or } 150$ beads) where the median Holm-adjusted p-value of the top 10 hotspots is lowest *and* where the mean “fraction interchromosomal” of the top 10 hotspots is greater than zero (“fraction interchromosomal” is the proportion of each hotspot that is not composed of the predominate chromosome). We started with $k=25$ as the smallest option since the 3D hotspots elicited for smaller values of k were mostly comprised of single genomic regions. Similarly, we conditioned that the selection of k must have a mean fraction interchromosomal of the top 10 hotspots greater than zero so that at least some interchromosomal 3D hotspots could be recovered.

PRIM seeks to identify hotspots by sequentially and strategically paring away data regions so that the average outcome over the remaining data is elevated. At each iteration, a fraction of the beads (*peel.alpha*) are peeled off the reconstruction by evaluating the extremal

slices orthogonal to each of the coordinate axes and removing data from whichever slice results in the highest mean ChIP-seq peak height for the remaining beads. This process is continued until a prescribed minimum number of beads (*min_beads*) remains. The resultant region can be enlarged to correct potential overshoot by pasting additional beads on to the region (via *paste.alpha*, which is smaller than *peel.alpha*) if that increases the mean ChIP-seq peak height (see “Methods”). At this point, a PRIM region or “box” has been identified. The beads comprising this box are then excluded and the entire procedure is repeated to identify additional boxes. We performed inference as follows: for each PRIM box, we preserved the beads comprising that box, then permuted ChIP-seq peak height values along each chromosome, and computed the mean ChIP-seq peak height of the box with the permuted data to generate a null referent distribution for estimating a p-value (see “Methods”). We tuned *min_beads* for PRIM using the same procedure as for tuning *k* for *k*-NN regression.

3D hotspots identified for Swi6

Swi6 is a component of two different transcription factor complexes: SBF (composed of Swi6 and the sequence-specific transcription factor Swi4) and MBF (composed of Swi6 and the sequence-specific transcription factor Mbp1)^{139,140}. SBF and MBF regulate genes that function in G1/S (e.g. cell growth genes, DNA synthesis genes)^{139,140}.

In accordance with our criterion we selected $k=50$ for *k*-NN regression of Swi6 peak height on the 3D genome reconstruction. Table 4.1 lists the top 10 resulting Swi6 3D hotspots, of which three were significant after multiple testing correction (see “Methods”). We focus on one of these significant Swi6 3D hotspots to illustrate the potential of such hotspot elicitation.

The 1st-ranked Swi6 3D hotspot from k -NN ($k=50$) regression contains a region from chromosome 7 (chr7) and a region from chr8 (Figure 4.2). Notably, each of these regions contains one of the 207 Swi6 target genes previously identified in a Swi6 ChIP-on-chip analysis¹³⁹. The first region contains the cell adhesion mucin gene MSB2¹⁴¹. The second region contains the ergosterol (cell membrane sterol) biosynthesis gene ERG11¹⁴².

Next, we applied PRIM ($min_beads=25$, selected in accordance with the tuning criterion above) to the 3D genome reconstruction overlaid with Swi6 peak height. Table 4.2 lists the top 10 resulting Swi6 3D hotspots, all of which were significant after multiple testing correction. Encouragingly, two of the genomic regions in the 1st-ranked (tied) Swi6 3D hotspot from PRIM (Figure 4.3) are the same as those in the 1st-ranked Swi6 3D hotspot from k -NN regression (MSB2 on chr7 and ERG11 on chr8).

Thus, k -NN regression and PRIM both identified a significant and corresponding Swi6 3D hotspot. This 3D hotspot is comprised of genomic regions from different chromosomes that contain known Swi6 target genes. As noted, identification of such regions absent a 3D reconstruction – in particular from the native interaction frequencies – is prohibitive due to the vastness of the attendant search space.

Rank	Beads	Genomic regions (by chromosome)	Mean ChIP-seq Peak Height	p-value (Holm)
1	50	chr 7;8	0.37	3.3e-02*
2	50	chr 4;4;4	0.36	3.3e-02*
3	50	chr 9;11	0.37	4.3e-02*
4	50	chr 4;12	0.37	6.9e-02
5	50	chr 15	0.39	8.9e-02
6	50	chr 4;4	0.38	8.9e-02
7	50	chr 7;8;8	0.42	8.9e-02
8	50	chr 15	0.40	8.9e-02
9	50	chr 4;12	0.43	9.5e-02
10	50	chr 2	0.44	1.5e-01

Table 4.1: The top 10 Swi6 3D hotspots from k -NN ($k=50$) regression

The top 10 hotspots (ranked by raw p-value). Holm-adjusted p-values are shown (*asterisks*: $p < 0.05$). Genomic regions comprising the hotspot are listed by chromosome. Regions from the same chromosome are listed separately when their gap is greater than 50 kilobases.

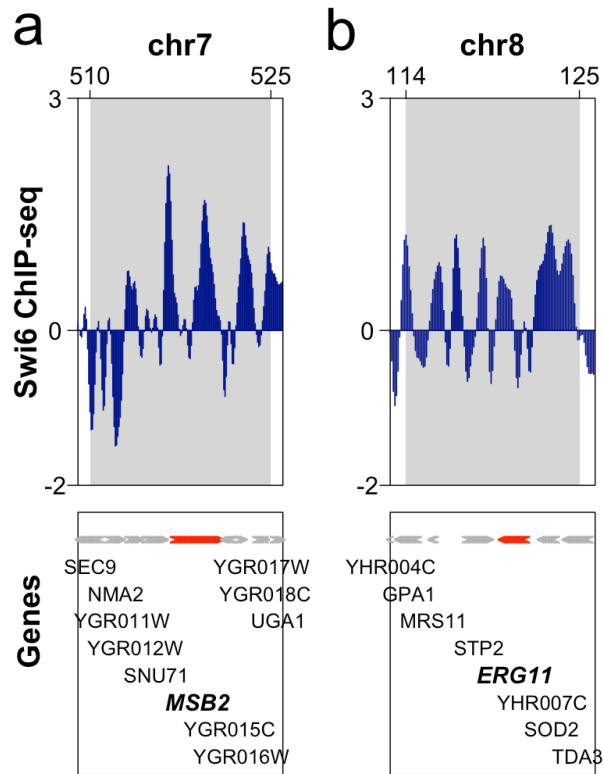


Figure 4.2: Genomic regions comprising the 1st-ranked Swi6 3D hotspot from k -NN ($k=50$) regression

This 3D hotspot contains one region from chr7 and one region from chr8. For each region, the top panel shows the \log_2 -normalized ChIP-seq peak height in that region (*gray* background) and in 1-kb of flanking sequence (*white* background). For each region, the bottom panel shows the genes in that region by their names, genomic positions, and orientations. The genes MSB2 and ERG11 (highlighted in *red*) were previously identified as significant Swi6 target genes in a ChIP-on-chip analysis¹³⁹.

Rank	Beads	Genomic regions (by chromosome)	Mean ChIP-seq Peak Height	p-value (Holm)
1	25	chr 7;8;16	0.82	6.4e-04*
1	29	chr 10;10;15	0.74	6.4e-04*
1	25	chr 12;15;15	0.92	6.4e-04*
1	31	chr 6;6;14	0.68	6.4e-04*
1	25	chr 12	0.87	6.4e-04*
6	31	chr 2;2;2;2	0.64	3.2e-03*
7	25	chr 14;14;14	0.69	1.1e-02*
8	30	chr 15;15;15	0.62	1.4e-02*
9	25	chr 1;2	0.66	1.6e-02*
10	28	chr 13;15	0.63	2.8e-02*

Table 4.2: The top 10 Swi6 3D hotspots from PRIM (*min_beads*=25)

The top 10 hotspots (ranked by raw p-value). Holm-adjusted p-values are shown (*asterisks*: $p < 0.05$). Genomic regions comprising the hotspot are listed by chromosome. Regions from the same chromosome are listed separately when their gap is greater than 50 kilobases.

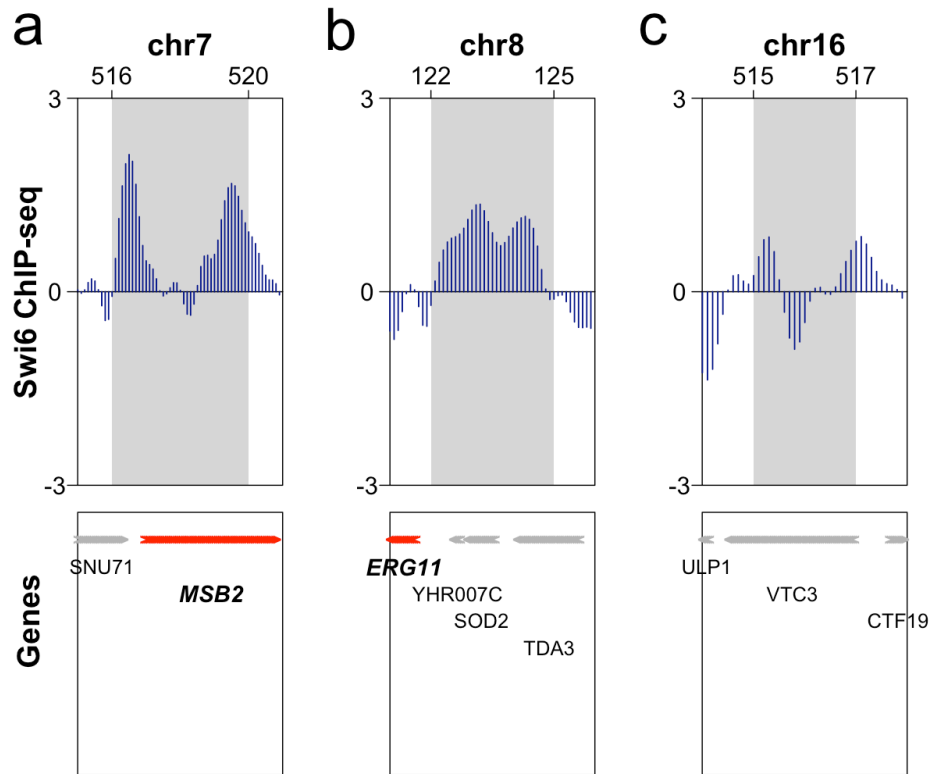


Figure 4.3: Genomic regions comprising the 1st-ranked Swi6 3D hotspot (chr 7;8;16) from PRIM (*min_beads*=25)

This 3D hotspot contains one region from chr7, one region from chr8, and one region from chr16. For each region, the top panel shows the log₂-normalized ChIP-seq peak height in that region (*gray* background) and in 1-kb of flanking sequence (*white* background). For each region, the bottom panel shows the genes in that region by their names, genomic positions, and orientations. The genes MSB2 and ERG11 (highlighted in *red*) were previously identified as significant Swi6 target genes in a ChIP-on-chip analysis¹³⁹.

3D hotspots identified for Pol2Ser5p

Pol2Ser5p is a marker of the active transcriptional machinery¹⁴³. We applied k -NN regression ($k=75$, as selected by our tuning criterion) to the 3D genome reconstruction overlaid with Pol2Ser5p peak height. Table 4.3 lists the top 10 resulting Pol2Ser5p 3D hotspots, all of which were significant after multiple testing correction. We showcase one of the most significant of these Pol2Ser5p 3D hotspots. This 3D hotspot contains three genomic regions from chr10 and one genomic region from chr13 (Figure 4.4). Notably, the latter three of these four regions contain genes that function in vacuole transport or vesicle transport^{126,144}: PEP8 (region 2)¹⁴⁵, VPS55 (region 3)¹⁴⁶, and TVP18 (region 4)¹⁴⁷ (Figure 4b-d). The second region (Figure 4b) also contains snR60¹⁴⁸, a non-coding RNA that functions in RNA binding/processing^{126,144}.

Next, we applied PRIM ($min_beads=75$, as selected by our tuning criterion) to the 3D genome reconstruction superposed with Pol2Ser5p peak height. Table 4.4 lists the top 10 resulting Pol2Ser5p 3D hotspots, of which seven were significant after multiple testing correction. Again there is concordance between methods with three of the genomic regions in the 1st-ranked Pol2Ser5p 3D hotspot from PRIM (Figure 4.5) being the same as those in the 1st-ranked (tied) Pol2Ser5p 3D hotspot from k -NN regression.

This top ranked Pol2Ser5p 3D hotspot from PRIM also contains multiple genes that function in vacuole transport or vesicle transport^{126,144} from different genomic regions: VPS55 (region 4)¹⁴⁶, TVP18 (region 5)¹⁴⁷, and ATG4 (region 6)¹⁴⁹. Additionally, it contains multiple genes that function in RNA binding/processing^{126,144} from different genomic regions: snR60 (region 2)¹⁴⁸, TMA22 (region 3)¹⁵⁰, SQS1 (alias PFA1; region 6)¹⁵¹, SSU72 (region 6)¹⁵², POP1 (region 6)¹⁵³, and NOP15 (region 7)¹⁵⁴.

Thus, k -NN regression and PRIM both identified a significant and corresponding Pol2Ser5p 3D hotspot. This 3D hotspot is comprised of distal genomic regions that contain functionally related genes.

Rank	Beads	Genomic regions (by chromosome)	Mean ChIP-seq Peak Height	p-value (Holm)
1	75	chr 10;10;10;13	0.44	8.6e-04*
1	75	chr 3;12;16;16	0.39	8.6e-04*
3	75	chr 15	0.49	1.2e-03*
4	75	chr 13;13;13	0.52	1.2e-03*
5	75	chr 15;15;15	0.39	1.3e-03*
6	75	chr 15;15;15	0.54	1.9e-03*
7	75	chr 15;15;15	0.44	1.9e-03*
8	75	chr 15;15;15;15	0.39	1.9e-03*
9	75	chr 13;13;13;13	0.40	1.9e-03*
10	75	chr 13;13;13;13	0.56	2.8e-03*

Table 4.3: The top 10 Pol2Ser5p 3D hotspots from k -NN ($k=75$) regression

The top 10 hotspots (ranked by raw p-value). Holm-adjusted p-values are shown (*asterisks*: $p < 0.05$). Genomic regions comprising the hotspot are listed by chromosome. Regions from the same chromosome are listed separately when their gap is greater than 50 kilobases.

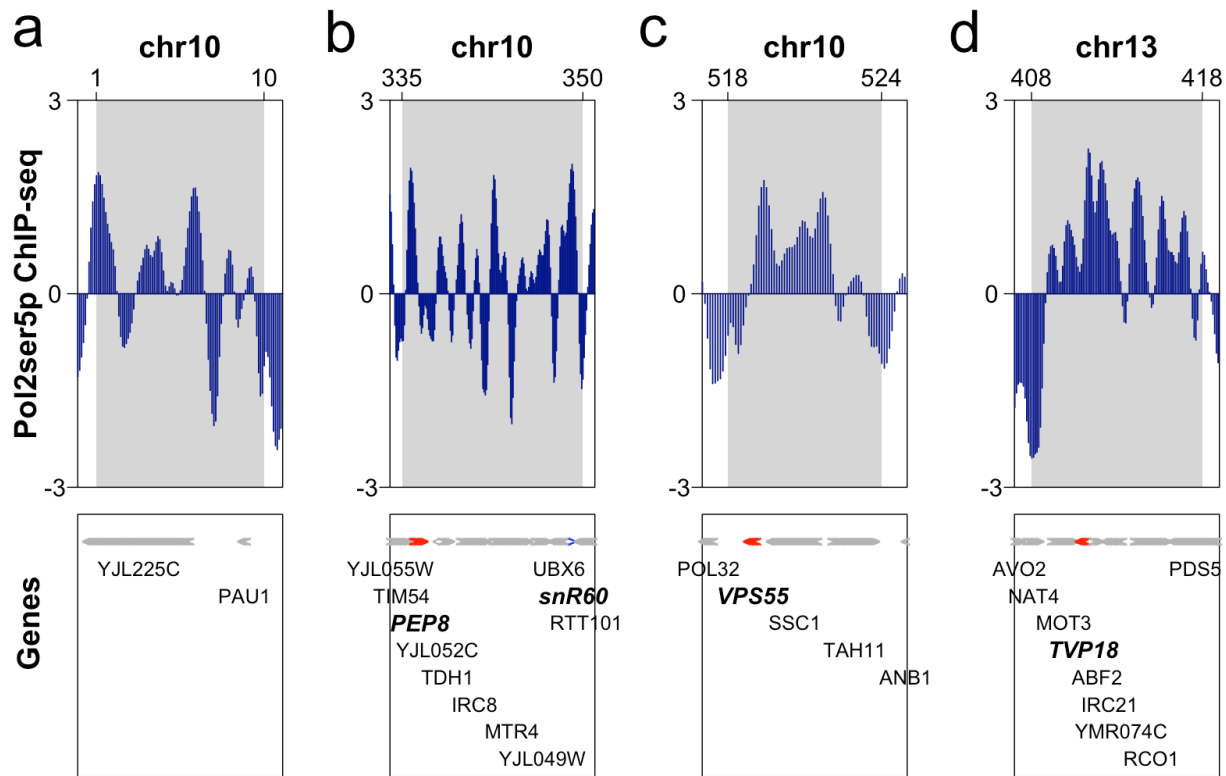


Figure 4.4: Genomic regions comprising the 1st-ranked Pol2Ser5p 3D hotspot (chr 10;10;10;13) from k -NN ($k=75$) regression

This 3D hotspot contains three regions from chr10 and one region from chr13. For each region, the top panel shows the \log_2 -normalized ChIP-seq peak height in that region (*gray* background) and in 1-kb of flanking sequence (*white* background). For each region, the bottom panel shows the genes in that region by their names, genomic positions, and orientations. The genes PEP8, VPS55, and TVP18 (highlighted in *red*) function in vacuole transport or vesicle transport. The gene snR60 (highlighted in *blue*) functions in RNA binding/processing.

Rank	Beads	Genomic regions (by chromosome)	Mean ChIP-seq Peak Height	p-value (Holm)
1	82	chr 10;10;10;10;13;14;14	0.65	2.3e-04*
2	77	chr 12;12	0.53	1.6e-03*
3	78	chr 9;9;15	0.49	4.2e-03*
4	78	chr 10;10;10	0.48	5.1e-03*
5	91	chr 12;15;15;16	0.45	6.4e-03*
6	75	chr 5;5;5;10;16	0.51	1.1e-02*
7	99	chr 2;7;8;13	0.37	2.4e-02*
8	76	chr 4;7;12;15;15;15;15	0.42	5.3e-02
9	75	chr 9;10;11;11	0.42	7.9e-02
10	75	chr 2;2;2;2	0.37	9.2e-02

Table 4.4: The top 10 Pol2Ser5p 3D hotspots from PRIM (*min_beads*=75)

The top 10 hotspots (ranked by raw p-value). Holm-adjusted p-values are shown (*asterisks*: $p < 0.05$). Genomic regions comprising the hotspot are listed by chromosome. Regions from the same chromosome are listed separately when their gap is greater than 50 kilobases.

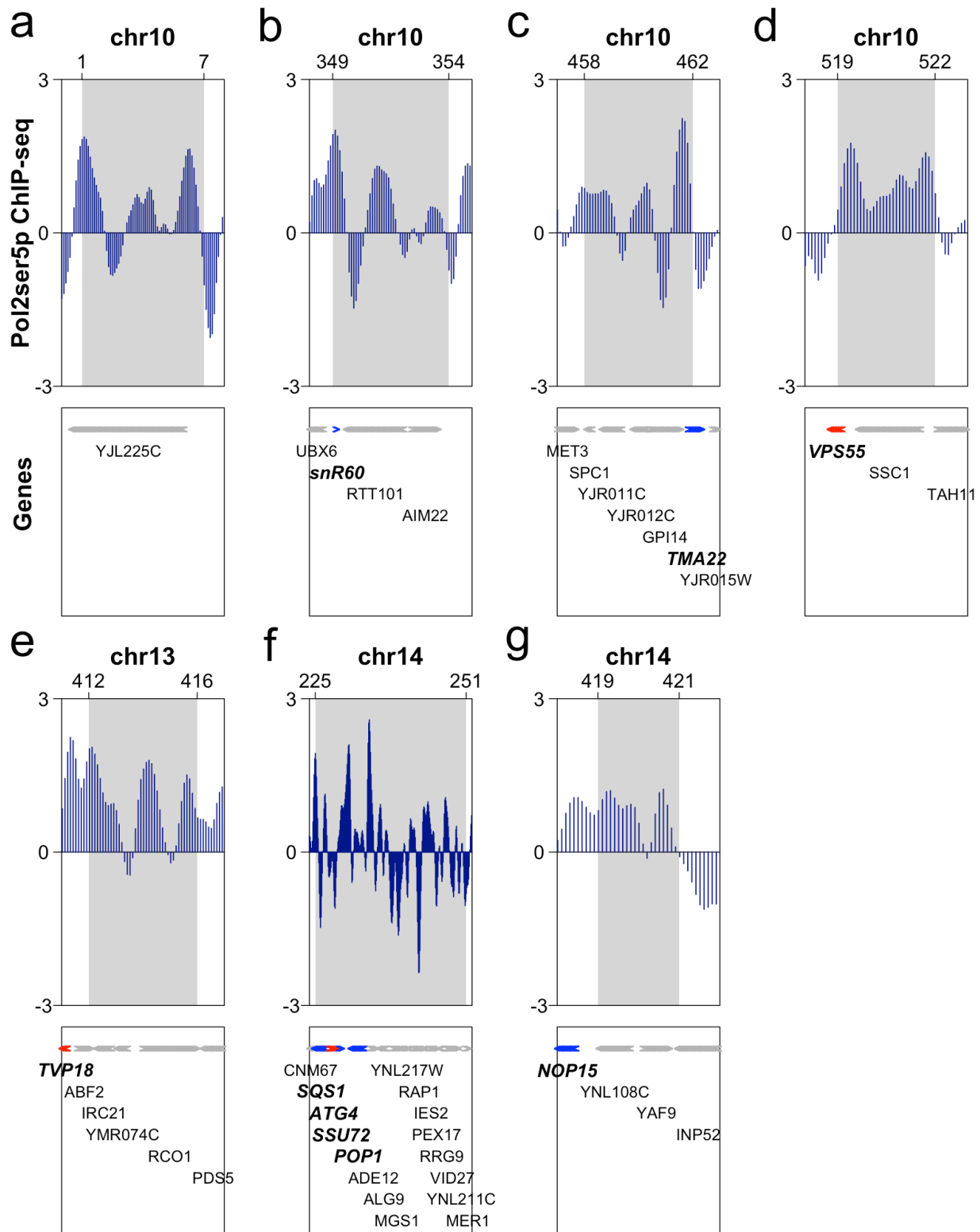


Figure 4.5: Genomic regions comprising the 1st-ranked Pol2Ser5p 3D hotspot from PRIM (*min_beads*=75; legend continued on next page)

This 3D hotspot contains four regions from chr10, one region from chr13, and two regions from chr14. For each region, the top panel shows the log₂-normalized ChIP-seq peak height in that region (*gray* background) and in 1-kb of flanking sequence (*white* background). For each region, the bottom panel shows the genes in that region by their names, genomic positions, and orientations. The genes VPS55, TVP18, and ATG4 (highlighted in *red*) function in vacuole transport or vesicle transport. The genes snR60, TMA22, SQS1, SSU72, POP1, and NOP15 (highlighted in *blue*) function in RNA binding/processing.

3D hotspots identified for Tup1

Tup1 is a transcriptional repressor that mediates glucose repression in *S. cerevisiae*¹⁵⁵. Tup1 also represses genes involved in hypoxia response, DNA damage response, and mating type switch¹⁵⁵. Tup1 does not bind DNA directly, but is recruited by several sequence-specific transcription factors¹⁵⁵.

We applied *k*-NN regression (with *k*=50 selected by our tuning criterion) to the 3D genome reconstruction overlaid with Tup1 peak height. Table 4.5 lists the top 10 resulting Tup1 3D hotspots, all of which were significant after multiple testing correction. We highlight one of these significant Tup1 3D hotspots that contains two distal regions from chr4 that separated by >400 kilobases. Notably, both of these regions contain one of the 149 genes that previous microarray analyses identified as significantly de-repressed in *tup1*^{E463A} mutants¹⁵⁶ (Figure 4.6). The first region contains the gene NRG1, which encodes a transcription factor that recruits Tup1¹⁵⁷ and that is glucose-repressed itself¹⁵⁸ perhaps via auto-regulation¹⁵⁸. The second region contains the gene YAP6, which also encodes a transcription factor that recruits Tup1¹⁵⁹ and for which similar auto-regulation would be plausible.

Next, we applied PRIM (with *min_beads*=75 selected by our tuning criterion) to the 3D genome reconstruction overlaid with Tup1 peak height. Table 4.6 lists the top 10 resulting Tup1 3D hotspots, of which nine were significant after multiple testing correction. Two of the genomic regions in the 4th-ranked Tup1 3D hotspot from PRIM (Figure 4.7) are the same as those in the 1st-ranked (tied) Tup1 3D hotspot from *k*-NN regression (containing NRG1 and YAP6).

Thus, *k*-NN regression and PRIM both identified a significant and corresponding Tup1 3D hotspot. This 3D hotspot is comprised of distal genomic regions that contain genes known to be regulated by Tup1.

Rank	Beads	Genomic regions	Mean ChIP-seq Peak Height	p-value (Holm)
1	50	chr 12;16	0.73	1.0e-05*
1	50	chr 15;15	0.43	1.0e-05*
1	50	chr 4;4	0.43	1.0e-05*
1	50	chr 15;15	0.43	1.0e-05*
1	50	chr 15;15;15;15	0.42	1.0e-05*
1	50	chr 12;12	0.41	1.0e-05*
7	50	chr 12;12	0.44	1.6e-05*
8	50	chr 5;15	0.45	1.3e-04*
9	50	chr 12;12;12	0.47	1.3e-04*
10	50	chr 12;16	0.47	2.8e-04*

Table 4.5: The top 10 Tup1 3D hotspots from *k*-NN (*k*=50) regression

The top 10 hotspots (ranked by raw p-value). Holm-adjusted p-values are shown (*asterisks*: $p < 0.05$). Genomic regions comprising the hotspot are listed by chromosome. Regions from the same chromosome are listed separately when their gap is greater than 50 kilobases.

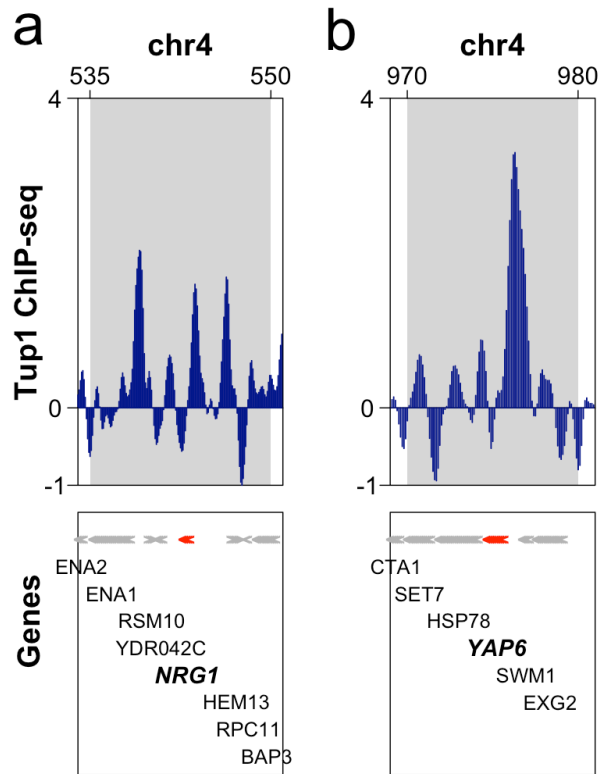


Figure 4.6: Genomic regions comprising the 1st ranked Tup1 3D hotspot (chr 4;4) from *k*-NN (*k*=50) regression

This 3D hotspot contains two regions from chr4. For each region, the top panel shows the log₂-normalized ChIP-seq peak height in that region (*gray* background) and in 1-kb of flanking sequence (*white* background). For each region, the bottom panel shows the genes in that region by their names, genomic positions, and orientations. The genes *NRG1* and *YAP6* (highlighted in *red*) were previously identified as significantly de-repressed in a microarray analysis of *tup1*-mutants¹⁵⁶.

Rank	Beads	Genomic regions	Mean ChIP-seq Peak Height	p-value (Holm)
1	81	chr 4;4;12;16	0.48	2.2e-04*
1	77	chr 4;10;12;12;16	0.45	2.2e-04*
1	77	chr 15;15;15;15	0.42	2.2e-04*
4	75	chr 4;4;10;16	0.37	4.4e-04*
5	77	chr 2;2;8;15	0.36	6.5e-04*
6	75	chr 15;15;15	0.35	4.8e-03*
7	78	chr 10;10;13;13;13;13	0.32	6.7e-03*
8	77	chr 2;7;7;7	0.29	1.3e-02*
9	78	chr 4;4	0.29	3.2e-02*
10	86	chr 4;4;10;10;12;16	0.26	7.2e-02

Table 4.6: The top 10 Tup1 3D hotspots from PRIM (*min_beads*=75)

The top 10 hotspots (ranked by raw p-value). Holm-adjusted p-values are shown (*asterisks*: $p < 0.05$). Genomic regions comprising the hotspot are listed by chromosome. Regions from the same chromosome are listed separately when their gap is greater than 50 kilobases.

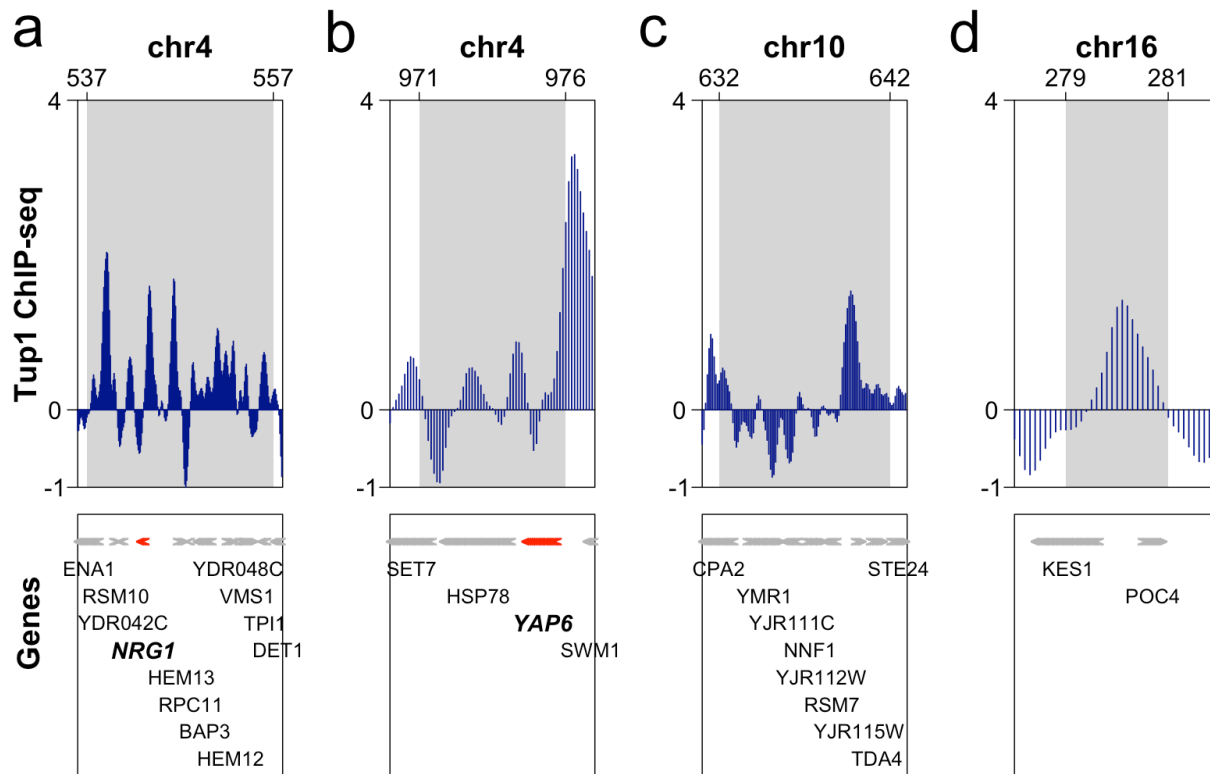


Figure 4.7: Genomic regions comprising the 4th ranked Tup1 3D hotspot from PRIM (*min_beads*=75)

This 3D hotspot contains two regions from chr4, one region from chr10, and one region from chr16. For each region, the top panel shows the log₂-normalized ChIP-seq peak height in that region (*gray* background) and in 1-kb of flanking sequence (*white* background). For each region, the bottom panel shows the genes in that region by their names, genomic positions, and orientations. The genes NRG1 and YAP6 (highlighted in *red*) were previously identified as significantly de-repressed in a microarray analysis of *tup1*-mutants¹⁵⁶.

Sensitivity of PRIM results to orientation of the 3D genome reconstruction

The results of k -NN regression are invariant to rotations of the 3D genome reconstruction, an important attribute given that the 3D genome reconstruction is coordinate-free: its orientation with respect to the X, Y, and Z axes is arbitrarily set. On the other hand, because PRIM operates by peeling off data points orthogonally to the X, Y, and Z axes at each iteration, its results depend on the (arbitrary) orientation of the 3D genome reconstruction. Accordingly, we assessed the sensitivity of the PRIM results to rotations of the 3D genome reconstruction.

We applied a rotation matrix to the original 3D genome reconstruction to generate six rotated 3D genome reconstructions: the three possible combinations of two-angle 45-degree rotations, and the three possible combinations of two-angle 315-degree rotations. We chose 45 and 315 degrees so that the resulting rotated 3D reconstructions would not be orthogonal to the original 3D reconstruction, and so that the outcome of this analysis would provide a representation of the orientation dependence of the original PRIM results.

We (separately) overlaid the six rotated 3D genome reconstructions with the ChIP-seq peak height for each input and then applied PRIM using the same setting of *min_beads* as previously. For each ChIP-seq input, we took the beads comprising the PRIM hotspot illustrated in the manuscript and then found the top rank by p-value of the PRIM hotspots from rotated 3D reconstructions in which these beads were present.

For Swi6, the top ranked PRIM hotspots from the six rotated 3D reconstructions in which beads from the original hotspot were present were ranked 9 (out of 650), 48 (out of 632), 29 (out of 667), 40 (out of 661), 10 (out of 685), and 1 (out of 615; Table 4.7). For Pol2Ser5p, the top ranked PRIM hotspots from the six rotated 3D reconstructions in which beads from the original hotspot were present were ranked 6 (out of 210), 1 (out of 218), 7 (out of 208), 16 (out of 205),

13 (out of 224), and 2 (out of 210; Table 4.8). For Tup1, the top ranked PRIM hotspots from the six rotated 3D reconstructions in which beads from the original hotspot were present were ranked 7 (out of 215), 2 (out of 212), 1 (out of 207), 1 (out of 210), 20 (out of 222), and 8 (out of 227; Table 4.9).

If we consider a rank ≤ 10 as an acceptable level of orientation dependence (i.e. there would at least be the opportunity for some of the same genomic regions to be recovered in downstream biological analyses), then these results indicate only modest orientation dependence of the PRIM results for Pol2ser5p and Tup1, with greater dependence for Swi6 although, even then, there is considerable stability of identified regions over the rotations (note: the results for Swi6 have a greater total number of PRIM boxes, since our tuning criterion selected a smaller value of *min_beads* for Swi6 than for Pol2Ser5p and Tup1).

Rotation <i>x, y, z</i>	Top Rank (Number of Boxes)
<i>0, 0, 0</i>	1 (out of 644)
<i>45, 45, 0</i>	9 (out of 650)
<i>45, 0, 45</i>	48 (out of 632)
<i>0, 45, 45</i>	29 (out of 667)
<i>315, 315, 0</i>	40 (out of 661)
<i>315, 0, 315</i>	10 (out of 685)
<i>0, 315, 315</i>	1 (out of 615)

Table 4.7: Top box-ranks of beads from the original Swi6 PRIM 3D hotspot (highlighted in the manuscript) when PRIM (*min_beads* = 25) is applied to rotated 3D reconstructions

Rotation <i>x, y, z</i>	Top Rank (Number of Boxes)
<i>0, 0, 0</i>	1 (out of 224)
<i>45, 45, 0</i>	6 (out of 210)
<i>45, 0, 45</i>	1 (out of 218)
<i>0, 45, 45</i>	7 (out of 208)
<i>315, 315, 0</i>	16 (out of 205)
<i>315, 0, 315</i>	13 (out of 224)
<i>0, 315, 315</i>	2 (out of 210)

Table 4.8: Top box-ranks of beads from the original Pol2Ser5p PRIM 3D hotspot (highlighted in the manuscript) when PRIM (*min_beads* = 75) is applied to rotated 3D reconstructions

Rotation <i>x, y, z</i>	Top Rank (Number of Boxes)
<i>0, 0, 0</i>	1 (out of 221)
<i>45, 45, 0</i>	7 (out of 215)
<i>45, 0, 45</i>	2 (out of 212)
<i>0, 45, 45</i>	1 (out of 207)
<i>315, 315, 0</i>	1 (out of 210)
<i>315, 0, 315</i>	20 (out of 222)
<i>0, 315, 315</i>	8 (out of 227)

Table 3.9: Top box-ranks of beads from the original Tup1 PRIM 3D hotspot (highlighted in the manuscript) when PRIM (*min_beads* = 75) is applied to rotated 3D reconstructions

Stability of k -NN regression and PRIM results over parameter settings

Though the results of PRIM depend on the 3D genome reconstruction's orientation (unlike the rotationally invariant k NN regression results), PRIM is more adaptive with respect to the extent of elicited 3D hotspots than k NN regression. The reason for this is that the choice of k specified for k NN regression is prescriptive: all resulting 3D hotspots will contain exactly k beads. Moreover, results corresponding to differing values of k must be obtained enumeratively: there is no means of updating findings from one value to another. Conversely, setting *min_beads* for PRIM only dictates the minimum number of beads in each 3D hotspot with expansion via the subsequent pasting steps. This putative recovery via pasting suggests starting with small settings of *min_beads* but on account of interplay with pasting and peeling parameters, along with computational considerations, exploration of *min_beads* on results is still warranted. We explored the sensitivity of the results to decreasing *paste.alpha* below the default setting (0.01): this had minimal impact of the results.

Accordingly, we assessed the stability of the downstream biological findings of the 3D hotspots illustrated in the manuscript over settings of k for k -NN regression and settings of *min_beads* for PRIM. For each ChIP-seq input and each algorithm, we determined which settings of k or *min_beads* (out of 25, 50, 75, 100, 125, and 150; see Figure 4.8) yielded a 3D hotspot that contains at least two of the same genes of interest (from at least two distinct genomic regions) as the 3D hotspot illustrated in the manuscript above for that ChIP-seq input. These results are depicted in Table 4.10.

For Swi6 with k -NN regression, $k=50$ was the only setting that identified a significant 3D hotspot containing the genes of interest MSB2 (chr7) and ERG11 (chr8). For Swi6 with PRIM, multiple settings of *min_beads* (25, 75, and 150) identified a significant 3D hotspot containing

these two genes. For Pol2Ser5p with k -NN regression, $k=75$ was the only setting that identified a significant 3D hotspot containing the genes highlighted previously from chr10, chr13, and chr14 that function in vacuole transport or RNA binding/processing. For Pol2Ser5p with PRIM, multiple settings of *min_beads* (75, 100, 125, and 150) identified a significant 3D hotspot containing at least two of these genes from distinct genomic regions. For Tup1 with k -NN regression, multiple settings of k (50, 100, and 150) identified a significant 3D hotspot containing the genes of interest NRG1 and YAP6. For Tup1 with PRIM, *min_beads*=75 was the only setting that identified a significant 3D hotspot containing these two genes. Thus, for two out of the three ChIP-seq inputs, the downstream biological findings of the 3D hotspot were more stable over parameter settings for PRIM than for k -NN regression.

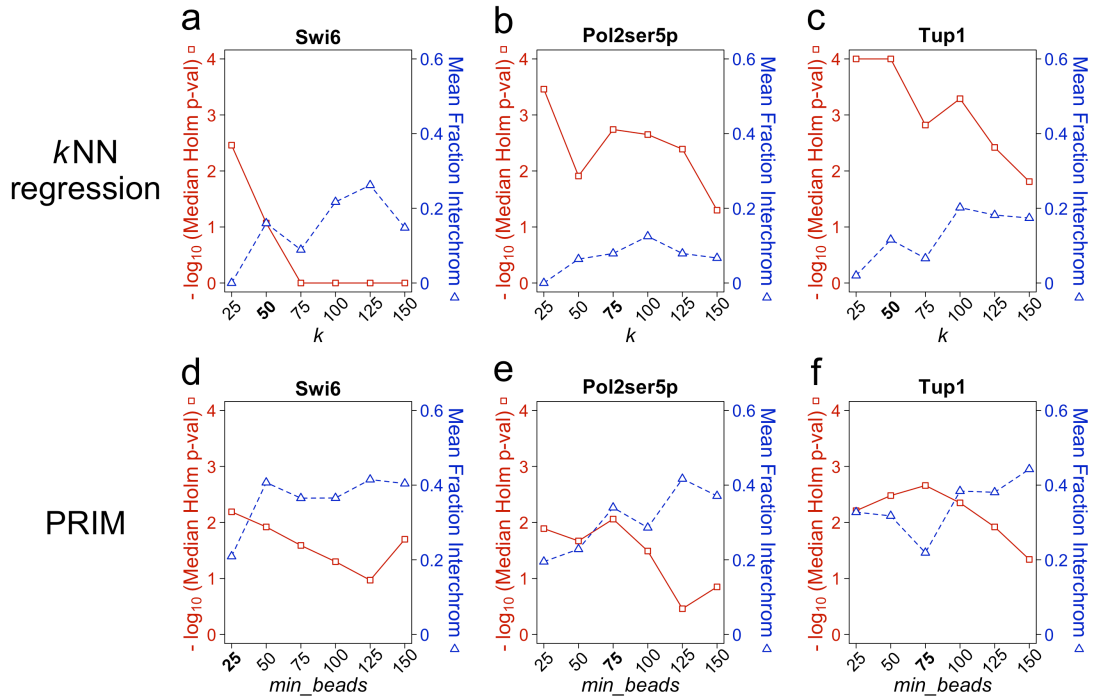


Figure 4.8: Parameter tuning for k -NN regression and PRIM

Top row: tuning k for k -NN regression. Bottom row: tuning min_beads for PRIM. The values of two variables are shown over a range of settings of k or min_beads . Left y-axis ($-\log_{10}$ scale): the median Holm-adjusted p-value of the top 10 hotspots (for 10^5 permutations). Right y-axis: the mean fraction interchromosomal of the top 10 hotspots (“fraction interchromosomal” is the proportion of each hotspot that is *not* composed of the predominant chromosome). The selected k or min_beads is shown in bold on the x-axis for each ChIP-seq input: (a) Swi6, (b) Pol2Ser5p, (c) Tup1.

		<i>k</i> or <i>min_beads</i>
Swi6	<i>k</i> -NN regression	50
	PRIM	25 , 75, 150
Pol2Ser5p	<i>k</i> -NN regression	75
	PRIM	75 , 100, 125, 150
Tup1	<i>k</i> -NN regression	50 , 100, 150
	PRIM	75

Table 4.10: Stability of downstream biological findings over settings of *k* or *min_beads*

For each ChIP-seq input and each algorithm, the setting of *k* or *min_beads* used in the manuscript to showcase downstream biological findings is shown in bold. Additional settings of *k* or *min_beads* (out of 25, 50, 75, 100, 125, 150) are listed if they yield the same downstream biological finding: i.e. a significant 3D hotspot containing at least two of the same genes of interest from at least two distinct genomic regions.

4.5 Conclusions

For each ChIP-seq input superposed on the 3D genome reconstruction, *k*-NN regression and PRIM identified a significant and corresponding 3D hotspot of ChIP-seq peak height. These 3D hotspots are comprised of regions that are either intra-chromosomally distal or from multiple chromosomes and contain known target genes of the transcriptional regulator (in the case of Swi6 and Tup1) or contain functionally related genes (in the case of Pol2Ser5p).

An important attribute of *k*-NN regression is that its results are invariant to rotations of the (coordinate-free) 3D genome reconstruction, while the results of PRIM depend on the orientation of the 3D reconstruction because it peels off beads orthogonally to the X, Y, and Z axes. On the other hand, PRIM is arguably more robust to parameter tuning than *k*-NN

regression because it performs a pasting step for each 3D hotspot identified: coupling this with prescription of a small value for the peeling parameter enables efficient and adaptive exploration of an extensive range of candidate hotspot solution regions. The orientation dependence of PRIM can be assessed by re-applying it to (disparate) rotated 3D genome reconstructions. It is of course possible to obtain rotational invariance by performing an initial rotation of the 3D genome coordinates to, say, its principal axes. However, the attendant PRIM solution is still dependent on the resultant coordinate reference frame.

Other techniques could potentially be used to identify 3D hotspots. Recursive partitioning or tree-structured regression methods⁵⁸ can isolate regions by successive splitting. However, it was partly to overcome the top-down greediness of these approaches that PRIM was advanced. Like PRIM these methods are not invariant under rotation but such invariance can be attained using splits that are linear combinations of the coordinate axes, but due to computational expensive and instability these methods are disfavored⁵⁸. Approaches based on algebraic topology and in particular persistent homology and Betti number barcodes¹⁶⁰ have possible utility in eliciting 3D hotspots but are undeveloped from an inferential perspective.

We have emphasized that effective identification of 3D hotspots is contingent on exploiting a 3D reconstruction and unattainable from searching combinations of raw (pairwise) contacts (interaction frequencies) due to combinatorial explosion therein. Conversely, hotspots so obtained are conditional on the 3D reconstruction used and gauging the accuracy or even the reproducibility thereof remains challenging¹²⁴. Accuracy assessments have made recourse to agreements with select FISH markers (e.g. ^{45,161}), while other approaches use FISH measurements to calibrate Hi-C derived distances¹⁶². However, there remains a sizeable disparity in the resolution of these data types and unless FISH markers happened to coincide with

emergent hotspots such evaluations may not be informative. The emergence of Hi-C assays for single cell data¹²⁰ allows for the possibility of a series of carefully designed experiments with associated 3D genome reconstructions to at least better address questions of reconstruction reproducibility. Even with such developments validation of putative 3D hotspots obtained as described herein, will require recourse to custom experimentation.

4.6 Methods

Hi-C contact data normalization and generating a 3D reconstruction

The *S. cerevisiae* Hi-C contact data (*Hind*III, pre-FDR, no masking) from⁴³ (Supplementary Data) was downloaded from <https://noble.gs.washington.edu/proj/yeast-architecture/sup.html>. We normalized this contact data for GC content, mappability, and fragment length by applying HiCNorm¹¹² genome-wide (chromosome by chromosome) as per¹³⁵. The HiCNorm source code was downloaded from <http://www.people.fas.harvard.edu/~junliu/HiCNorm/> (last update: “08.05.2012”). We filtered by interaction frequency to retain the top contacts and then generated a new 3D reconstruction from this normalized and filtered contact data using the constrained optimization approach from⁴³.

ChIP-seq data normalization

The raw ChIP-seq dataset from¹³⁶, which contains three input samples (Swi6, Pol2Ser5p, and Tup1) and a mock immunoprecipitation (IP) control sample (DMSO, Illumina), was obtained from GEO (dataset GSE51251; <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51251>). We converted the raw sequencing SRA data to FASTQ format using “fastq-dump” version 2.3.4 in the Sequence Read

Archive (SRA) Toolkit¹⁶³. We aligned the reads to the *S. cerevisiae* reference genome (sacCer2) using Bowtie 2¹⁶⁴ version 2.2.1 with default parameters and then converted the SAM output to BAM format using SAMtools¹⁶⁵ version 0.1.19-44428cd. We filtered the sequencing reads (using the R package “ShortRead”¹⁶⁶ version 1.20.0 with a custom filter) to retain only those with 2 or less expected errors per read: given a Phred quality score q for each base call, the probability that a base call is erroneous is $p = 10^{-(q/10)}$, which is then summed over the bases in the read to give the expected errors per read. We deduplicated the reads (using “ShortRead”¹⁶⁶) to control for PCR amplification bias^{50,51}. We masked the highly repetitive chr12 rDNA region (using “ShortRead”¹⁶⁶ with a custom filter) because of the difficulty of aligning short reads there. We log₂-normalized each ChIP-seq sample to the mock IP control using the function `get.smoothed.enrichment.mle()` in the R package “spp”¹⁶⁷ version 1.11 with a 200 basepair (bp) bandwidth and 100 bp stepsize. “spp” was downloaded from <http://compbio.med.harvard.edu/Supplements/ChIP-seq/>.

We performed a smoothing step to control for the local dependency of the signal in linear genomic space, since we are interested in identifying 3D hotspots of physically proximal yet genomically distal ChIP-seq peaks in subsequent analyses. Specifically, we smoothed each normalized ChIP-seq signal along each chromosome arm using SuperSmoother¹⁶⁸, which is implemented as `supsmu()` in R (version 3.0.2) package “stats”, with the span determined by cross-validation. We then took the residuals of each smoothed normalized signal — this constitutes the final “ChIP-seq peak height” that we proceeded to superpose onto the 3D chromatin configuration reconstruction.

Superposing ChIP-seq peak height on the 3D reconstruction

The 3D chromatin configuration reconstruction consists of a series of beads spaced along each chromosome in the genome; each bead has a genomic position and (X,Y,Z) coordinates. For each ChIP-seq input (Swi6, Pol2Ser5p, Tup1), we binned its peak height data (i.e. the residuals of the smoothed log₂-normalized signal) such that each bin was centered on a bead. We then assigned to each bead the most extreme ChIP-seq peak height (positive or negative) from the bin centered on that bead. The result is a 3D chromatin configuration reconstruction overlaid with functional genomic data: each bead now has a genomic position, physical coordinates (X,Y,Z), and a ChIP-seq peak height value. We visualized this superposed 3D reconstruction in MacPyMOL 1.3¹⁶⁹ by first converting the data to the Protein Data Bank (PDB) file format¹⁷⁰ with the ChIP-seq peak height value rescaled as the temperature factor (B-factor) in the PDB file.

k-Nearest Neighbor (k-NN) regression

We applied *k*-NN regression using the R package “FNN”¹⁷¹ to identify hotspots in ChIP-seq peak height superposed on the 3D reconstruction. *k*-NN regression is performed as follows. For each bead in the superposed 3D genome reconstruction, the function `knn.index()` returns the row indices of the *k* beads closest to the bead in Euclidean distance, and the function `knn.reg()` returns the mean ChIP-seq peak height of the beads corresponding to that *k*-NN group (the input bead and its *k* nearest neighbor beads). The *k*-NN groups are then ranked by mean ChIP-seq peak height, and the top 10 groups are retained for statistical inference (more below).

Rather than applying these functions to every bead in the superposed 3D reconstruction, we applied them to “seed” beads evenly spaced along each chromosome by `seed_interval = 0.2*k`. The reason for this is to reduce redundancy in the top 10 *k*-NN groups and to allow

distinct 3D hotspots to be elicited (otherwise, the top 10 k -NN groups would primarily consist of genomically adjacent beads all corresponding to a single 3D hotspot).

Once the top 10 k -NN groups by mean ChIP-seq peak height have been identified, significance is then estimated by permutation. The ChIP-seq peak height values are permuted along each chromosome, k -NN regression is re-applied to the permuted data, and the top 10 resulting k -NN groups are saved. This process is repeated for a total of 10^6 permutations. P-values are estimated by comparing the mean ChIP-seq peak height of the k -NN groups from the observed data to the mean ChIP-seq peak height values of the k -NN groups from the permuted data *along each rank*. For example, the p-value of the *third* ranked k -NN group is estimated by comparing its mean ChIP-seq peak height to the mean ChIP-seq peak height values of the *third* ranked k -NN groups across all of the permutations. We Holm-adjusted the p-values for multiple testing.

Patient Rule Induction Method (PRIM)

We applied PRIM using the R package “prim”⁵⁷ to identify hotspots in ChIP-seq peak height superposed on the 3D reconstruction. We applied `prim.box()` to the data (using the default settings `peel.alpha=0.05` and `paste.alpha=0.01`). This returns statistics on each of the boxes identified (e.g. the number of beads in the box; the mean ChIP-seq peak height of the box; the (X,Y,Z) boundaries of the box). We then applied `predict()` to the output of `prim.box()` plus the original data, which returns the mapping of each bead to the appropriate PRIM box label (the PRIM boxes are numerically labeled; the largest numerical label is a placeholder for the beads that were not boxed).

The superposed 3D reconstruction now has an additional column: each bead has a genomic position, physical (X,Y,Z) coordinates, a ChIP-seq peak height value, and a PRIM box number. We estimated the significance of all of the PRIM boxes (except for the placeholder) by permutations as follows. We preserved the mapping of beads to PRIM boxes, and then permuted the ChIP-seq peak height along each chromosome. Then we computed the mean ChIP-seq peak height for each PRIM box from the permuted data. We repeated this for a total of 10^6 permutations. The p-value of each box was estimated by comparing its mean ChIP-seq peak height from the observed data to its mean ChIP-seq peak height values from the permuted data. We Holm-adjusted the p-values for multiple testing.

4.7 Acknowledgments

Some computations were performed using the UCSF Biostatistics High Performance Computing System. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1144247 and National Institutes of Health Grant R01 GM109457. DC was supported in part by the National Institutes of Health Training Grant T32 GM007175.

Chapter 5: Discussion

Throughout this dissertation, we have demonstrated that applying appropriate data preprocessing in conjunction with advanced supervised learning algorithms improves the interpretability of next-generation sequencing studies of chromatin structure and organization. Specifically, we focused on the preprocessing and analysis of histone modification ChIP-seq data (Chapter 2), of 3D genome reconstructions from Hi-C data (Chapter 3), or of both combined (histone modification ChIP-seq data superposed on 3D genome reconstructions; Chapter 4). In Chapter 2, we improved the preprocessing of ChIP-seq data compared to a previous study⁴⁹ (which used raw reads) by deduplicating reads to control for PCR amplification bias^{50,51}, down-sampling reads to control for variable sequencing depth⁵¹, and identifying stable nucleosome positions with NPS⁴⁴. In Chapter 3, we improved the preprocessing of the Hi-C contact data compared to a previous study⁵⁵ (which preceded the formalization of pipelines for redressing biases in Hi-C data^{108,111,112}) by correcting for fragment length, GC content, and mappability with HiCNorm¹¹². In Chapter 4, we similarly performed careful preprocessing of both the ChIP-seq data and Hi-C contact data before generating a 3D reconstruction and superposing ChIP-seq peak height onto the 3D reconstruction. A recent study applied clustering techniques to the contact pairs from Hi-C data based on epigenetic data¹⁷²; however, this was not a focal analysis and it did not make use of 3D reconstructions.

In addition to improved data preprocessing, our analyses involved the application of advanced algorithms that emphasize downstream interpretation. In Chapter 2, we performed feature selection for classification by applying DMFS⁵², which (in contrast to enumerative feature generation, e.g., all k -mers), avoids the generation of abundant noise features and allows longer, potentially informative sequence features to be tested. We performed classification by applying Random Forests⁵⁴, which, in contrast to the C4.5 (or any single) decision tree, has

likely predictive performance gains resulting from being an ensemble classifier, while still emphasizing interpretation by allowing the ranking of features by importance. In Chapter 3, we assessed the 3D localization of functional annotations by testing the significance of a test statistic: the Median of Pairwise Euclidean Distances (MPED). This was an improvement over previous analyses⁴⁵, which assessed the enrichment of “close” pairs (following dichotomization of the data) and resulted in significance levels that varied by dichotomization threshold. In Chapter 4, we adapted and applied to methodologies to analyze input data where each data point has a physical (X,Y,Z) coordinate paired with outcome value (ChIP-seq peak height) with the goal of identifying regions in 3-space for which the mean outcome is significantly elevated. Few if any existing methodologies seemed suited to this task. Our adaptations of *k*-NN regression and PRIM identified a significant and corresponding 3D hotspot for each ChIP-seq input analyzed.

As far as specific biological findings, in Chapter 2, we identified a significant and biologically meaningful DNA sequence feature associated with H2A/H4R3me2s: “TCCATT”, which is part of the consensus sequence of satellite II and III DNA^{74,75}. This finding is consistent with a recently discovered biochemical mechanism: H4R3me2s provides a binding site for the DNA methyltransferase (Dnmt3a)⁷⁰, which methylates satellite II and III DNA⁸⁵⁻⁸⁷. Appropriate data preprocessing was crucial to this discovery, as was the use a classification algorithm that allowed downstream ranking of the importance of individual features: Random Forests. Subsequent to our analyses, several studies have also employed Random Forests for other supervised learning / classification problems related to chromatin structure, for example: predicting chromatin boundaries from histone modification ChIP-seq data and Hi-C data¹⁷³; predicting transcription factor binding from genome-wide, nucleotide-resolution DNA methylation data¹⁷⁴; and predicting differentially expressed genes in lung cancer (based on RNA-

seq analysis of lung cancer and adjacent healthy tissue) from DNA methylation data and histone modification ChIP-seq data¹⁷⁵.

In addition, we demonstrated the added value of performing downstream biological analyses on 3D genome reconstructions rather than just on the Hi-C contact data. In Chapter 3, we detected sets of developmentally regulated genes in *P. falciparum* as significantly co-localized with reconstruction-based assessment but not with contact-based assessment (the latter being inherently limited to detecting strictly pairwise interactions). In Chapter 4, we identified significant 3D hotspots in *S. cerevisiae* ChIP-seq peak height superposed on a 3D genome reconstruction, which were composed of multiple (two to seven) distal genomic regions. For Swi6, we identified a 3D hotspot containing two known Swi6 target genes on different chromosomes. For Pol2ser5p, we identified a 3D hotspot containing multiple genes that function in vacuole transport and RNA binding/processing. For Tup1, we identified a 3D hotspot containing two known Tup1-regulated genes from distal regions of chromosome 4. Having a 3D reconstruction to guide such analyses is crucial to discovering such multi-region functional nuclear hotspots; identifying them solely from the (pairwise) Hi-C contact matrix would be computationally prohibitive. Nevertheless, realizing these advantages of performing downstream biological analyses on 3D genome reconstructions depends on the ability to generate a 3D reconstruction that is both accurate and consistent across experimental replicates (or from different restriction enzyme libraries). Statistical methods for gauging the consistency of 3D reconstructions have recently been advanced¹²⁴, as have methods to improve the accuracy of the Hi-C data and 3D reconstructions by calibration with FISH data¹⁶².

The field of Hi-C data generation and analysis is progressing very rapidly and there are several notable recent advances. Many analyses to date, including ours, are based on interphase

cells of two model organisms that are haploid and have relatively small genomes (compared to human): *S. cerevisiae* and *P. falciparum*. Previously, the resolution of human Hi-C data had been relatively low because the larger genome size requires much greater sequencing depth. Nevertheless, a recent study obtained kilobase-resolution Hi-C data for nine human cell types by generating over five terabytes of sequence data¹⁷⁶. Another challenge of mammalian Hi-C data analysis was the presence of diploid genomes; however, methods for inferring haplotypes from Hi-C data have been advanced¹²¹. Hi-C analysis of mitotic cells has recently been reported¹⁷⁷, as has the first example of single cell Hi-C analysis¹²⁰. Recent studies have also reinforced the value of Hi-C analysis to human disease research. Specifically, two studies applied Capture Hi-C to genomic regions that were significant in Genome-Wide Association Studies of diseases (but that had no nearby candidate genes) in order to identify distal interactions involving these regions^{178,179}.

Beyond the specific findings reported in this dissertation, the analytic methods that we advanced can be applied to myriad other high-throughput data sets of chromatin structure and 3D organization by researchers in the future. The approach that we employed for classifying histone modifications – DMFS followed by Random Forests – could readily be applied to the ChIP-seq data available for dozens of other histone modifications to gain insight into their biological properties and functions. In addition, our adaption and application of k -NN regression and PRIM to identify 3D hotspots of ChIP-seq peak height superposed on a 3D genome reconstruction could be applied to any type of continuous functional genomic data (e.g. other ChIP-seq data or RNA-seq data) superposed on a 3D genome reconstruction.

References

1. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet.* 2008;9:387–402.
2. Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science.* 2008;320(5881):1344–1349.
3. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature.* 2007;447(7146):799–816.
4. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57–74.
5. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. *Nature.* 2010;467(7319):1061–1073.
6. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012;491(7422):56–65.
7. Davuluri RV, Grosse I, Zhang MQ. Computational identification of promoters and first exons in the human genome. *Nat Genet.* 2001;29(4):412–417.
8. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* 1997;268(1):78–94.
9. Claverie JM. Computational methods for the identification of genes in vertebrate genomic sequences. *Human Molecular Genetics.* 1997;6(10):1735–1744.
10. Sonnenburg S, Schweikert G, Philips P, Behr J, Rätsch G. Accurate splice site prediction using support vector machines. *BMC Bioinformatics.* 2007;8 Suppl 10:S7.
11. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research.* 2004;32:D91–4.
12. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics.* 2011;27(7):1017–1018.
13. Erwin GD, Oksenberg N, Truty RM, et al. Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput Biol.* 2014;10(6):e1003677.
14. Ng KLS, Mishra SK. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics.* 2007;23(11):1321–1330.

15. Bell O, Tiwari VK, Thomä NH, Schübeler D. Determinants and dynamics of genome accessibility. *Nat Rev Genet.* 2011;12(8):554–564.
16. Misteli T. Beyond the sequence: cellular organization of genome function. *Cell.* 2007;128(4):787–800.
17. Kouzarides T, Berger S. Chromatin modifications and their mechanism of action. In: Allis C, Jenuwein T, Reinberg D, eds. *Epigenetics*. 1st ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2007:191–209.
18. Rando OJ. Global patterns of histone modifications. *Curr Opin Genet Dev.* 2007;17(2):94–99.
19. Xu X, Hoang S, Mayo MW, Bekiranov S. Application of machine learning methods to histone methylation ChIP-Seq data reveals H4R3me2 globally represses gene expression. *BMC Bioinformatics.* 2010;11:396.
20. Jenuwein T, Allis CD. Translating the histone code. *Science.* 2001;293(5532):1074–1080.
21. Elgin SCR, Reuter G. Position-effect variegation, heterochromatin formation, and gene silencing in *Drosophila*. *Cold Spring Harb Perspect Biol.* 2013;5(8):a017780.
22. Fodor BD, Shukeir N, Reuter G, Jenuwein T. Mammalian Su(var) genes in chromatin control. *Annu Rev Cell Dev Biol.* 2010;26:471–501.
23. Herold M, Bartkuhn M, Renkawitz R. CTCF: insights into insulator function during development. *Development.* 2012;139(6):1045–1057.
24. Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. *Nat Rev Genet.* 2007;8(4):253–262.
25. Badeaux AI, Shi Y. Emerging roles for chromatin as a signal integration and storage platform. *Nat Rev Mol Cell Biol.* 2013;14(4):211–224.
26. Fraga MF, Ballestar E, Paz MF, et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci USA.* 2005;102(30):10604–10609.
27. Sugawara H, Iwamoto K, Bundo M, et al. Hypermethylation of serotonin transporter gene in bipolar disorder detected by epigenome analysis of discordant monozygotic twins. *Transl Psychiatry.* 2011;1:e24.
28. Häslér R, Feng Z, Bäckdahl L, et al. A functional methylome map of ulcerative colitis. *Genome Res.* 2012;22(11):2130–2137.
29. Davies MN, Krause L, Bell JT, et al. Hypermethylation in the ZBTB20 gene is associated with major depressive disorder. *Genome Biol.* 2014;15(4):R56.

30. Yuan W, Xia Y, Bell CG, et al. An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins. *Nat Commun*. 2014;5:5719.
31. Portela A, Esteller M. Epigenetic modifications and human disease. *Nat Biotechnol*. 2010;28(10):1057–1068.
32. Ivanov M, Kacevska M, Ingelman-Sundberg M. Epigenomics and interindividual differences in drug response. *Clin Pharmacol Ther*. 2012;92(6):727–736.
33. Esteller M, Garcia-Foncillas J, Andion E, et al. Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N Engl J Med*. 2000;343(19):1350–1354.
34. Agrelo R, Cheng W-H, Setien F, et al. Epigenetic inactivation of the premature aging Werner syndrome gene in human cancer. *Proc Natl Acad Sci USA*. 2006;103(23):8822–8827.
35. Zimmer C, Fabre E. Principles of chromosomal organization: lessons from yeast. *J Cell Biol*. 2011;192(5):723–733.
36. Meister P, Taddei A. Building silent compartments at the nuclear periphery: a recurrent theme. *Curr Opin Genet Dev*. 2013;23(2):96–103.
37. Andrulis ED, Neiman AM, Zappulla DC, Sternglanz R. Perinuclear localization of chromatin facilitates transcriptional silencing. *Nature*. 1998;394(6693):592–595.
38. Shumaker DK, Dechat T, Kohlmaier A, et al. Mutant nuclear lamin A leads to progressive alterations of epigenetic control in premature aging. *Proc Natl Acad Sci USA*. 2006;103(23):8703–8708.
39. Misteli T. Higher-order genome organization in human disease. *Cold Spring Harb Perspect Biol*. 2010;2(8):a000794.
40. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer*. 2007;7(4):233–245.
41. Robertson G, Hirst M, Bainbridge M, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*. 2007;4(8):651–657.
42. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326(5950):289–293.
43. Duan Z, Andronescu M, Schutz K, et al. A three-dimensional model of the yeast genome. *Nature*. 2010;465(7296):363–367.
44. Zhang Y, Shin H, Song JS, Lei Y, Liu XS. Identifying positioned nucleosomes with

- epigenetic marks in human from ChIP-Seq. *BMC Genomics*. 2008;9(1):537.
45. Ay F, Bunnik EM, Varoquaux N, et al. Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res*. 2014;24(6):974–988.
 46. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. *Cell*. 2007;129(4):823–837.
 47. Wang Z, Zang C, Rosenfeld JA, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*. 2008;40(7):897–903.
 48. Henikoff S, Shilatifard A. Histone modification: cause or cog? *Trends Genet*. 2011;27(10):389–396.
 49. Gervais AL, Gaudreau L. Discriminating nucleosomes containing histone H2A.Z or H2A based on genetic and epigenetic information. *BMC Mol Biol*. 2009;10(1):18.
 50. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods*. 2009;6(11 Suppl):S22–32.
 51. Leleu M, Lefebvre G, Rougemont J. Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. *Briefings in Functional Genomics and Proteomics*. 2011;9(5-6):466–476.
 52. Xiong H, Capurso D, Sen S, Segal MR. Sequence-based classification using discriminatory motif feature selection. *PLoS ONE*. 2011;6(11):e27382.
 53. Magnan CN, Randall A, Baldi P. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics*. 2009;25(17):2200–2207.
 54. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5–32.
 55. Witten DM, Noble WS. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Research*. 2012;40(9):3849–3855.
 56. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*. 1992;46(3):175–185.
 57. Duong T. *Prim: Patient Rule Induction Method (PRIM)*. R package version 1.0.15; 2014.
 58. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd ed. New York, NY: Springer; 2009.
 59. Asbury TM, Mitman M, Tang J, Zheng WJ. Genome3D: a viewer-model framework for integrating and visualizing multi-scale epigenomic information within a three-

- dimensional genome. *BMC Bioinformatics*. 2010;11:444.
60. Allis C, Jenuwein T, Reinberg D. Overview and concepts. In: Allis C, Jenuwein T, Reinberg D, eds. *Epigenetics*. 1st ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2007:23–61.
 61. Lee W, Tillo D, Bray N, et al. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet*. 2007;39(10):1235–1244.
 62. Segal MR. Re-cracking the nucleosome positioning code. *Statistical Applications in Genetics and Molecular Biology*. 2008;7(1):Article14.
 63. Miele V, Vaillant C, d'Aubenton-Carafa Y, Thermes C, Grange T. DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Research*. 2008;36(11):3746–3756.
 64. Tillo D, Hughes TR. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics*. 2009;10:442.
 65. Ancelin K, Lange UC, Hajkova P, et al. Blimp1 associates with Prmt5 and directs histone arginine methylation in mouse germ cells. *Nat Cell Biol*. 2006;8(6):623–630.
 66. Song J, Fisher D. Nucleosome positioning in promoters: significance and open questions. In: Appasani K, ed. *Epigenomics: From Chromatin Biology to Therapeutics*. 1st ed. Cambridge: Cambridge University Press; 2012.
 67. Tolstorukov MY, Volfovsky N, Stephens RM, Park PJ. Impact of chromatin structure on sequence variability in the human genome. *Nat Struct Mol Biol*. 2011;18(4):510–515.
 68. Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet*. 2009;10(3):161–172.
 69. Liang K, Keles S. Normalization of ChIP-Seq data with control. *BMC Bioinformatics*. 2012;13:199.
 70. Zhao Q, Rank G, Tan YT, et al. PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing. *Nat Struct Mol Biol*. 2009;16(3):304–311.
 71. Blagus R, Lusa L. Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics*. 2010;11:523.
 72. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*. Boca Raton, FL: CRC; 1984.
 73. Bonasio R, Tu S, Reinberg D. Molecular signals of epigenetic states. *Science*. 2010;330(6004):612–616.

74. Eymery A, Callanan M, Vourc'h C. The secret message of heterochromatin: new insights into the mechanisms and function of centromeric and pericentric repeat sequence transcription. *Int J Dev Biol.* 2009;53(2-3):259–268.
75. Prosser J, Frommer M, Paul C, Vincent PC. Sequence relationships of three human satellite DNAs. *J Mol Biol.* 1986;187(2):145–155.
76. Kaplan N, Hughes TR, Lieb JD, Widom J, Segal E. Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology. *Genome Biol.* 2010;11(11):140.
77. Fernández M, Miranda-Saavedra D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic Acids Research.* 2012;40(10):e77.
78. Beck D, Brandl MB, Boelen L, Unnikrishnan A, Pimanda JE, Wong JWH. Signal analysis for genome wide maps of histone modifications measured by ChIP-seq. *Bioinformatics.* 2012;28(8):1062–1069.
79. Leslie C, Eskin E, Noble WS. The spectrum kernel: a string kernel for SVM protein classification. *Pac Symp Biocomput.* 2002:564–575.
80. Leslie C, Kuang R. Fast string kernels using inexact matching for protein sequences. 2004;5:1435–1455.
81. Ratsch G, Sonnenburg S, Scholkopf B. RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics.* 2005;21 Suppl 1:i369–77.
82. Sonnenburg S, Zien A, Philips P, Rätsch G. POIMs: positional oligomer importance matrices--understanding support vector machine-based signal detectors. *Bioinformatics.* 2008;24(13):i6–14.
83. Schultheiss SJ, Busch W, Lohmann JU, Kohlbacher O, Rätsch G. KIRMES: kernel-based identification of regulatory modules in euchromatic sequences. *BMC Bioinformatics.* 2009;10(Suppl 13):O1.
84. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics.* 2009;25(15):1952–1958.
85. Chen T, Tsujimoto N, Li E. The PWWP Domain of Dnmt3a and Dnmt3b Is Required for Directing DNA Methylation to the Major Satellite Repeats at Pericentric Heterochromatin. *Molecular and Cellular Biology.* 2004;24(20):9048–9058.
86. Oka M, Rodić N, Graddy J, Chang L-J, Terada N. CpG sites preferentially methylated by Dnmt3a in vivo. *J Biol Chem.* 2006;281(15):9901–9908.
87. Fanelli M, Caprodossi S, Ricci-Vitiani L, et al. Loss of pericentromeric DNA

- methylation pattern in human glioblastoma is associated with altered DNA methyltransferases expression and involves the stem cell compartment. *Oncogene*. 2008;27(3):358–365.
88. Jurkowska RZ, Jurkowski TP, Jeltsch A. Structure and function of mammalian DNA methyltransferases. *Chembiochem*. 2011;12(2):206–222.
 89. Rank G, Cerruti L, Simpson RJ, Moritz RL, Jane SM, Zhao Q. Identification of a PRMT5-dependent repressor complex linked to silencing of human fetal globin gene expression. *Blood*. 2010;116(9):1585–1592.
 90. Erukashvily NI, Donev R, Waisertreiger ISR, Podgornaya OI. Human chromosome 1 satellite 3 DNA is decondensed, demethylated and transcribed in senescent cells and in A431 epithelial carcinoma cells. *Cytogenet Genome Res*. 2007;118(1):42–54.
 91. Ting DT, Lipson D, Paul S, et al. Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science*. 2011;331(6017):593–596.
 92. Tsuda H, Takarabe T, Kanai Y, Fukutomi T, Hirohashi S. Correlation of DNA hypomethylation at pericentromeric heterochromatin regions of chromosomes 16 and 1 with histological features and chromosomal abnormalities of human breast carcinomas. *Am J Pathol*. 2002;161(3):859–866.
 93. Ringrose L, Paro R. Polycomb/Trithorax response elements and epigenetic memory of cell identity. *Development*. 2007;134(2):223–232.
 94. Liaw A, Wiener M. Classification and regression by randomForest. *R news*. 2002;2(3):18–22.
 95. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. *R package version 16*. 2011.
 96. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCr: Visualizing the performance of scoring classifiers. *R package version 10-4*. 2009.
 97. Wang G, Yu T, Zhang W. WordSpy: identifying transcription factor binding motifs by building a dictionary and learning a grammar. *Nucleic Acids Research*. 2005;33(Web Server issue):W412–6.
 98. Wang G, Zhang W. A steganalysis-based approach to comprehensive identification and characterization of functional regulatory elements. *Genome Biol*. 2006;7(6):R49.
 99. Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*. 2004;32(Database issue):D493–6.
 100. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. 2005;110(1-4):462–467.

101. de Wit E, de Laat W. A decade of 3C technologies: insights into nuclear organization. *Genes & Development*. 2012;26(1):11–24.
102. Marti-Renom MA, Mirny LA. Bridging the Resolution Gap in Structural Modeling of 3D Genome Organization. Bourne PE, ed. *PLoS Comput Biol*. 2011;7(7):e1002125.
103. Razin SV, Gavrillov AA, Pichugin A, Lipinski M, Iarovaia OV, Vassetzky YS. Transcription factories in the context of the nuclear and genome organization. *Nucleic Acids Research*. 2011;39(21):9085–9092.
104. Paulsen J, Lien TG, Sandve GK, et al. Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucleic Acids Research*. 2013;41(10):5164–5174.
105. Kruse K, Sewitz S, Babu MM. A complex network framework for unbiased statistical analyses of DNA-DNA contact maps. *Nucleic Acids Research*. 2013;41(2):701–710.
106. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2002;64(3):479–498.
107. Le Roch KG, Zhou Y, Blair PL, et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*. 2003;301(5639):1503–1508.
108. Imakaev M, Fudenberg G, McCord RP, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*. 2012;9(10):999–1003.
109. Freitas-Junior LH, Bottius E, Pirrit LA, et al. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature*. 2000;407(6807):1018–1022.
110. Scherf A, Figueiredo LM, Freitas-Junior LH. Plasmodium telomeres: a pathogen's perspective. *Curr Opin Microbiol*. 2001;4(4):409–414.
111. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*. 2011;43(11):1059–1065.
112. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*. 2012;28(23):3131–3133.
113. Tanizawa H, Iwasaki O, Tanaka A, et al. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Research*. 2010;38(22):8164–8177.
114. Jin QW, Fuchs J, Loidl J. Centromere clustering is a major determinant of yeast interphase nuclear organization. *J Cell Sci*. 2000;113 (Pt 11):1903–1912.
115. Tanaka A, Tanizawa H, Sriswasdi S, et al. Epigenetic Regulation of Condensin-

- Mediated Genome Organization during the Cell Cycle and upon DNA Damage through Histone H3 Lysine 56 Acetylation. *Molecular Cell*. 2012;48(4):532–546.
116. Gotta M, Laroche T, Formenton A, Maillet L, Scherthan H, Gasser SM. The clustering of telomeres and colocalization with Rap1, Sir3, and Sir4 proteins in wild-type *Saccharomyces cerevisiae*. *J Cell Biol*. 1996;134(6):1349–1363.
 117. Schober H, Kalck V, Vega-Palas MA, et al. Controlled exchange of chromosomal arms reveals principles driving telomere interactions in yeast. *Genome Res*. 2008;18(2):261–271.
 118. Thompson M, Haeusler RA, Good PD, Engelke DR. Nucleolar clustering of dispersed tRNA genes. *Science*. 2003;302(5649):1399–1401.
 119. Frey BJ, Dueck D. Clustering by passing messages between data points. *Science*. 2007;315(5814):972–976.
 120. Nagano T, Lubling Y, Stevens TJ, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*. 2013;502(7469):59–64.
 121. Selvaraj S, Dixon JR, Bansal V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol*. 2013;31(12):1111–1118.
 122. Varoquaux N, Ay F, Noble WS, Vert JP. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*. 2014;30(12):i26–i33.
 123. Zhang Z, Li G, Toh K-C, Sung W-K. 3D chromosome modeling with semi-definite programming and Hi-C data. *J Comput Biol*. 2013;20(11):831–846.
 124. Segal MR, Xiong H, Capurso D, Vazquez M, Arsuaga J. Reproducibility of 3D chromatin configuration reconstructions. *Biostatistics*. 2014;15(3):442–456.
 125. Di Rienzi SC, Collingwood D, Raghuraman MK, Brewer BJ. Fragile genomic sites are associated with origins of replication. *Genome Biol Evol*. 2009;1:350–363.
 126. Consortium GO. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;25(1):25–29.
 127. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*. 1998;9(12):3273–3297.
 128. Hinrichs AS, Karolchik D, Baertsch R, et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*. 2006;34(Database issue):D590–8.
 129. Byrne KP, Wolfe KH. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*. 2005;15(10):1456–1461.

130. Yip KY, Cheng C, Bhardwaj N, et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* 2012;13(9):R48.
131. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science.* 2002;295(5558):1306–1311.
132. Van Bortle K, Corces VG. Nuclear organization and genome function. *Annu Rev Cell Dev Biol.* 2012;28:163–187.
133. Taddei A, Gasser SM. Structure and function in the budding yeast nucleus. *Genetics.* 2012;192(1):107–129.
134. Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J. 3D genome reconstruction from chromosomal contacts. *Nat Methods.* 2014;11(11):1141–1143.
135. Capurso D, Segal MR. Distance-based assessment of the localization of functional annotations in 3D genome reconstructions. *BMC Genomics.* 2014;15:992.
136. Park D, Lee Y, Bhupindersingh G, Iyer VR. Widespread misinterpretable ChIP-seq bias in yeast. *PLoS ONE.* 2013;8(12):e83506.
137. Friedman JH, Fisher NI. Bump hunting in high-dimensional data. *Statistics and Computing.* 1999;9(2):123–143.
138. Li W, Gong K, Li Q, Alber F, Zhou XJ. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics.* 2015;31(6):960–962.
139. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature.* 2001;409(6819):533–538.
140. Horak CE, Luscombe NM, Qian J, et al. Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes & Development.* 2002;16(23):3017–3033.
141. Vadaie N, Dionne H, Akajagbor DS, Nickerson SR, Krysan DJ, Cullen PJ. Cleavage of the signaling mucin Msb2 by the aspartyl protease Yps1 is required for MAPK activation in yeast. *J Cell Biol.* 2008;181(7):1073–1081.
142. Turi TG, Loper JC. Multiple regulatory elements control expression of the gene encoding the *Saccharomyces cerevisiae* cytochrome P450, lanosterol 14 alpha-demethylase (ERG11). *J Biol Chem.* 1992;267(3):2046–2056.
143. Egloff S, Murphy S. Cracking the RNA polymerase II CTD code. *Trends Genet.* 2008;24(6):280–288.
144. Cherry JM, Hong EL, Amundsen C, et al. *Saccharomyces* Genome Database: the

- genomics resource of budding yeast. *Nucleic Acids Research*. 2012;40(Database issue):D700–5.
145. Bachhawat AK, Suhan J, Jones EW. The yeast homolog of H 58, a mouse gene essential for embryogenesis, performs a role in the delivery of proteins to the vacuole. *Genes & Development*. 1994;8(12):1379–1387.
 146. Schluter C, Lam KKY, Brumm J, et al. Global analysis of yeast endosomal transport identifies the vps55/68 sorting complex. *Mol Biol Cell*. 2008;19(4):1282–1294.
 147. Inadome H, Noda Y, Kamimura Y, Adachi H, Yoda K. Tvp38, Tvp23, Tvp18 and Tvp15: novel membrane proteins in the Tlg2-containing Golgi/endosome compartments of *Saccharomyces cerevisiae*. *Experimental Cell Research*. 2007;313(4):688–697.
 148. Lowe TM, Eddy SR. A computational screen for methylation guide snoRNAs in yeast. *Science*. 1999;283(5405):1168–1171.
 149. Kaufmann A, Beier V, Franquelim HG, Wollert T. Molecular mechanism of autophagic membrane-scaffold assembly and disassembly. *Cell*. 2014;156(3):469–481.
 150. Fleischer TC, Weaver CM, McAfee KJ, Jennings JL, Link AJ. Systematic identification and functional screens of uncharacterized proteins associated with eukaryotic ribosomal complexes. *Genes & Development*. 2006;20(10):1294–1307.
 151. Lebaron S, Froment C, Fromont-Racine M, et al. The splicing ATPase prp43p is a component of multiple preribosomal particles. *Molecular and Cellular Biology*. 2005;25(21):9269–9282.
 152. He X, Khan AU, Cheng H, Pappas DL, Hampsey M, Moore CL. Functional interactions between the transcription and mRNA 3' end processing machineries mediated by Ssu72 and Sub1. *Genes & Development*. 2003;17(8):1030–1042.
 153. Lygerou Z, Mitchell P, Petfalski E, Séraphin B, Tollervey D. The POP1 gene encodes a protein component common to the RNase MRP and RNase P ribonucleoproteins. *Genes & Development*. 1994;8(12):1423–1433.
 154. Granneman S, Petfalski E, Tollervey D. A cluster of ribosome synthesis factors regulate pre-rRNA folding and 5.8S rRNA maturation by the Rat1 exonuclease. *EMBO J*. 2011;30(19):4006–4019.
 155. Malavé TM, Dent SYR. Transcriptional repression by Tup1-Ssn6. *Biochem Cell Biol*. 2006;84(4):437–443.
 156. Green SR, Johnson AD. Genome-wide analysis of the functions of a conserved surface on the corepressor Tup1. *Mol Biol Cell*. 2005;16(6):2605–2613.
 157. Park SH, Koh SS, Chun JH, Hwang HJ, Kang HS. Nrg1 is a transcriptional repressor for glucose repression of STA1 gene expression in *Saccharomyces cerevisiae*. *Molecular*

- and Cellular Biology*. 1999;19(3):2044–2050.
158. Berkey CD, Vyas VK, Carlson M. Nrg1 and nrg2 transcriptional repressors are differently regulated in response to carbon source. *Eukaryotic Cell*. 2004;3(2):311–317.
 159. Hanlon SE, Rizzo JM, Tatomer DC, Lieb JD, Buck MJ. The stress response factors Yap6, Cin5, Phd1, and Skn7 direct targeting of the conserved co-repressor Tup1-Ssn6 in *S. cerevisiae*. *PLoS ONE*. 2011;6(4):e19060.
 160. Adler RJ, Bobrowski O, Borman MS, Subag E, Weinberger S. Persistent homology for random fields and complexes. *Institute of Mathematical Statistics Collections, Volume 6*. 2010:124–143.
 161. Hu M, Deng K, Qin Z, et al. Bayesian inference of spatial organizations of chromosomes. *PLoS Comput Biol*. 2013;9(1):e1002893.
 162. Shavit Y, Hamey FK, Lio P. FisHiCal: an R package for iterative FISH-based calibration of Hi-C data. *Bioinformatics*. 2014;30(21):3120–3122.
 163. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Research*. 2011;39(Database issue):D19–21.
 164. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–359.
 165. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079.
 166. Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*. 2009;25(19):2607–2608.
 167. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*. 2008;26(12):1351–1359.
 168. Friedman JH. A variable span smoother. *Tech Rep Stanford LCS 5*. 1984:1–32.
 169. The PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC.
 170. Bernstein FC, Koetzle TF, Williams GJ, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol*. 1977;112(3):535–542.
 171. Beygelzimer A, Kakadet S, Langford J, Arya S, Mount D, Li S. *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. R package version 1.1; 2013.
 172. Lan X, Witt H, Katsumura K, et al. Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. *Nucleic Acids Research*. 2012;40(16):7690–7704.

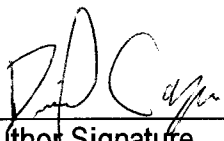
173. Bednarz P, Wilczyński B. Supervised learning method for predicting chromatin boundary associated insulator elements. *J Bioinform Comput Biol*. 2014;12(6):1442006.
174. Xu T, Li B, Zhao M, et al. Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Research*. 2015;43(5):2757–2766.
175. Li J, Ching T, Huang S, Garmire LX. Using epigenomics data to predict gene expression in lung cancer. *BMC Bioinformatics*. 2015;16 Suppl 5:S10.
176. Rao SSP, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014;159(7):1665–1680.
177. Naumova N, Imakaev M, Fudenberg G, et al. Organization of the mitotic chromosome. *Science*. 2013;342(6161):948–953.
178. Dryden NH, Broome LR, Dudbridge F, et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res*. 2014;24(11):1854–1868.
179. Jäger R, Migliorini G, Henrion M, et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun*. 2015;6:6178.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

6/9/2015

Date