UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Analysis of 3D genome organization and gene regulation in mammalian cells

Permalink

https://escholarship.org/uc/item/9j9818h0

Author

Selvaraj, Siddarth Gautham

Publication Date

2014

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Analysis of 3D genome organization and gene regulation in mammalian cells

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

in

Bioinformatics and System Biology

by

Siddarth Gautham Selvaraj

Committee in charge:

Professor Bing Ren, Chair
Professor Vineet Bafna, Co-Chair
Professor Alexander Hoffman
Professor Wei Wang
Professor Kun Zhang

2014

The Dissertation of Siddarth Gautham Selvaraj is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2014

DEDICATION

To ammama, appa, amma, tinku, and gulfi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My dissertation has benefitted from collaborations with several members of Ren lab. In particular, I would like to thank Jesse Dixon for generating most of the datasets that I analyzed throughout my graduate work, for his mentorship and friendship. I also would like to thank Gary Hon and Celso Espinoza for their wonderful support, and for always taking the time and effort to bring the better in me. I also enjoyed interactions with Andrea, Anu, Fulai, Ah-Young, Haruhiko, Nisha, Inkyung, Anthony, Danny, Dave, and other Ren lab members during this time. I am also extremely thankful to Alex, Roy, Chris, Andy, Teddy and other Bioinformatics program friends for all the formal and informal chats.

I am indebted to my thesis advisory committee members Vineet Bafna, Kun Zhang, Wei Wang and Alex Hoffman for their guidance and useful suggestions. Specifically, I would like to thank and appreciate Vineet for funding my first year of graduate school, in spite of not having a chance to work with him previously and yet letting me to join a lab of my choice for my thesis work. Finally, I am extremely grateful to my advisor Bing Ren, whose mentorship and calm demeanor have helped me immensely to grow as an individual.

Chapter 2, in full, is a reprint of the material published in Nature 2012. Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, Bing Ren. “Topological domains in mammalian genomes identified by analysis of chromatin interactions”, Nature 485 (7398), 376-380, 2012. The dissertation author was a primary investigator and author of this paper. In

particular, the dissertation author developed a computational method to identify topological domains and performed analyses to characterize these domains, such as their stability among cell-types and their conservation across species.

Chapter 3, in full, is a reprint of the material published in Nature Biotechnology 2013. Siddarth Selvaraj*, Jesse R Dixon*, Vikas Bansal, Bing Ren. “Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing”, Nature Biotechnology 31 (12), 1111-1118. 2013. The dissertation author was a primary investigator and author of this paper. Specifically, the dissertation author performed all of the computational analyses described in the paper.

Chapter 4, in full, has been submitted for review in Nature. Jesse R Dixon*, Inkyung Jung*, Siddarth Selvaraj*, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Victor V Lobanenkov, Joseph Ecker, James Thomson, Bing Ren. “Global Reorganization of Chromatin Architecture during Embryonic Stem Cell Differentiation”. The dissertation author was a primary investigator and author of this paper. In particular, the dissertation author performed haplotyping of H1 cells and analyzed gene regulation patterns in an allele-specific manner.

VITA

2006 Bachelor of Technology in Industrial Biotechnology, Anna University, Chennai

2009 Professional Master of Science in Computational Bioscience, Arizona State University, Tempe

2014 Doctor of Philosophy in Bioinformatics and Systems Biology, University of California, San Diego

PUBLICATIONS

Dixon JR*, Jung I*, Selvaraj S*, Shen Y, Antosiewica-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, Lobanekov V, Ecker J, Thomson J, Ren B. Global reorganization of chromatin architecture during embryonic stem cell differentiaton. Submitted to Nature.

Selvaraj S*, Dixon JR*, Bansal V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nature biotechnology. 2013; 31(12),1111-1118.

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

Hu, M., Deng, K., Qin, Z., Dixon, J., Selvaraj, S., Fang, J., Ren, B. & Liu, J.S. Bayesian inference of spatial organizations of chromosomes. PLoS computational biology 9, e1002893 (2013).

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):3131-3.

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

ABSTRACT OF THE DISSERTATION

Analysis of 3D genome organization and gene regulation in mammalian cells

by

Siddarth Gautham Selvaraj

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2014

Professor Bing Ren, Chair

Professor Vineet Bafna, Co-Chair

The three-dimensional structure of the genome plays a key role in gene regulation. For example, while highly compacted heterochromatin drives gene silencing, open euchromatin facilitates gene activation. Nevertheless, how chromatin folds within these structures and consequently how it controls access to genomic content is poorly understood. Recent advances in high-throughput sequencing have provided valuable tools, such as Hi-C, for the study of

chromatin structure. Using Hi-C datasets, I developed a hidden markov model based algorithm to identify self-interacting patterns of chromatin structure termed topological domains. These mega-base sized domains are pervasive through the genome and are highly conserved among human and mouse.

At a higher resolution, topological domains encompass individual chromatin interactions between regulatory elements and its target gene. Therefore, in order to mechanistically understand gene regulation, it is essential to elucidate the functional relationship among regulatory elements and their target genes. By exploiting the sequence diversity between homologous chromosomes, it is possible to delineate this relationship. However, this requires the knowledge of haplotypes, which has traditionally been difficult to obtain. As the Hi-C protocol preferentially recovers DNA variants on the same chromosome, I invented HaploSeq to reconstruct chromosome-scale haplotypes. HaploSeq can generate haplotypes with ~99.5% accuracy for >95% of alleles in mouse and 98% accuracy for ~81% of alleles in humans, thus solving a long-standing problem in genetics.

By integrating the knowledge of haplotypes, we queried the relationship between regulatory elements and gene expression in human embryonic stem cells and a panel of differentiated cell-types. Across the 5 cell lineages examined, I identified a total of 24% of genes that showed allelic bias in gene expression. While most of the allelic-genes had a correlating allelic-promoter chromatin state, ~29% of genes were exceptions suggesting other mechanisms of gene regulation. Accordingly, I then analyzed histone-acetylation marks to identify

1589 allelic enhancers. By predicting chromatin interactions using Hi-C, we observed allelic enhancers to be spatially proximal to allelic genes, suggesting cooperative activity among genome sequence, structure, and function.

Taken together, our studies suggest that gene regulation is facilitated and coordinated by genome structure.

Chapter 1: Interplay between genome structure and gene regulation

Abstract

Conventional genome sequencing technologies utilize pools of genomic DNA, which are fragmented prior to sequencing, resulting in the loss of three-dimensional (3D) genome information. The 3D genome offers critical insights into how cells interpret genetic and epigenetic content, and therefore is key for a mechanistic understanding of genome regulation. For example, precise control of transcription involves physical structural interactions among genes and distal regulatory elements. Recent advancements in molecular biology techniques and corresponding computational methods have allowed for accurate measurements of 3D genome structure, enabling targeted and genome-wide analyses of higher-order chromatin structure. Here, I review our current understanding of genome structure and it's utility in unraveling multiple aspects of genome regulation.

Introduction

The human genome project determined the genetic sequence that constitutes the human DNA[1-4], but how cells read, interpret, and control this information is less clear. Differences in deciphering genetic content can lead to variable gene regulation and transcription patterns, resulting in hundreds of unique cell-types and potentially numerous disease states in the human body[5-8]. Therefore, a fine-level understanding of the mechanisms behind gene regulation is critical for delineating the role of genetics in human health and disease.

In eukaryotes, gene regulation requires combinatorial functional activities involving regulatory elements such as promoters, non-coding RNAs, enhancers, and silencers[9]. To this end, the Roadmap Epigenome[10-15] and the ENCODE[16,17] consortiums have generated comprehensive profiles of DNA methylation, histone modifications, chromatin accessibility, and transcription-factor (TF) binding, allowing systematic annotation of regulatory elements. However, how these elements cooperate in a combinatorial fashion to facilitate gene regulation is poorly understood. As the eukaryotic genome is organized in non-random three-dimensional structures, knowledge of the 3D genome can reveal physical connections among genes and regulatory elements and thereby can further our understanding of gene regulation[18-20].

Recent technological advancements have allowed for measurement of 3D genome at different resolutions. For instance, while Fluorescence in-situ hybridization (FISH) has revealed patterns of chromosome territorial organization, chromosome conformation capture (C-technologies) has allowed chromatin structure studies of specific gene loci[20, 21]. Each of these studies have been valuable in showing the role of genome structure in its function. In addition, 3D genome information has been shown to be useful for deconvoluting chromosome-scale haplotype patterns[22]. Therefore, by exploiting aspects of 3D genome structural information, we can learn novel mechanistic aspects and potentially build predictive models of human gene regulation in a haplotype-resolved context.

Higher-order chromatin structures facilitate gene regulation

In the interphase of a eukaryotic cell's nucleus, the genome is non-randomly organized at multiple levels[18, 23]. For example, several FISH[21, 24, 25] and live cellular imaging[26, 27] based studies have revealed that chromosomes occupy distinct territories of nuclear positioning, termed chromosome territories (CTs). Further, several independent methods have indicated the physical and functional separation of active (euchromatin) and inactive (heterochromatin) regions of CTs[18, 19, 23]. On the one hand, active regions within a CT are generally positioned at the border of the resident CT and can interact with active regions from other CTs to allow co-regulation of genes[28]. On the other, independent methylation measurements of cells treated with Dam protein[29,30] and ChIP-Seq measurement of H3k9me3 histone tails[31] have demonstrated that inactive regions are physically associated with structures at the nuclear periphery, largely separated from the active regions. Such differential positioning of active and inactive regions allows for efficient usage of cellular machinery and agrees well with the transcription factory model of genome regulation[28]. Therefore, nuclear positioning of chromosomes and their ability to intermingle with each other and other nuclear structures has profound impact on global transcription.

While microscopy, live imaging, and ChIP-Seq studies have demonstrated aspects of genome positioning at the nuclear level, a higher resolution picture of specific structures within chromosome territories are lacking. Recent advancements in chromosome conformation capture (3C)[32-35] based methods have allowed us to investigate genome structure at the level of genes. In brief,

3C based methods work by crosslinking cells to retain the 3D chromatin structure. Then, the chromatin is fragmented and the crosslinked fragments are ligated to form new artificial fragments, which are then PCR amplified and/or sequenced. As 3C based methods generate fragment interaction frequencies, the spatial distance between the fragments and consequently genome structure can be delineated[32]. However, as 3C based methods are often performed on million of cells, each with dynamic 3D genome structures and at different cell-cycle phases, robust computational methods that can understand the stochasticity and true biological variability in the data have to be developed to generate meaningful 3D structure predictions.

Utilizing a variant of 3C, called the Hi-C[35], Job Dekker and colleagues profiled the genome-wide chromatin interaction patterns to observe two distinct compartments within CTs. These results correlated well with previously established active and inactive positioning of chromosomal regions[28, 29]. As this study lacked the sequencing depth to investigate chromosome structures at higher resolution, we performed Hi-C in human and mouse cells with ultra-deep sequencing to identify pervasive structural units of chromosomes termed Topological domains, or Topological associated domains (TADs)[36]. TADs are structures within the active and inactive compartments. We used rigorous non-parametric computational methods to remove systematic biases in Hi-C data owing to variability across fragments in terms of fragment length, GC content and its mappability[37]. Then, we implemented a hidden markov model that predicted TAD locations in the genome with high confidence. TADs are megabase-sized

domains of high local chromatin interaction frequency yet well spatially separated from other TADs. In addition, intervening boundary sequences between TADs are invariant among cell-types and conserved between human and mouse. More recently, TADs have also been identified in drosophila[38], demonstrating an evolutionary aspect of genome structure.

The topological domain-like organization of chromosomes is well established in the literature[23,39,40]. In particular, FISH and 3C based studies have revealed correlation between changes in domain structure and gene regulation[41-43]. We have also shown evidence that suggest TADs can constrain chromatin interactions between genes and regulatory elements and such intra-TAD interactions are more involved in cell-type specific gene regulation patterns[36,44]. In addition, we have revealed that TAD boundaries correspond to insulator activity of transcription and that the boundaries correlate well with structural transition events that mark several functional activities – such as replication timing, and specification of inactive regions that move towards the nuclear periphery[36]. Recent restraint based iterative modeling of chromatin interaction data has allowed building of sophisticated 3D conformations of TADs and their relative positioning in a chromosome[45, 46]. In addition, 3D modeling of HoxA and α-globin domains has illustrated the dynamics of chromatin structure and gene expression across a panel of cell-types[41, 43, 46]. Undoubtedly, identification of TADs and modeling of their conformations have enabled systematic analyses of chromatin structure at a resolution that reveals dynamic localization of group of genes.

While chromatin structures such as TADs and CTs seem to be static across a population of cells, structures measured at a deeper resolution have revealed that interactions among genomic loci can be dynamic. In particular, the single-cell Hi-C[47] study revealed structural stochasticity at the gene level but consistent intermingling patterns of active domains of several CTs. Similarly, Jin and colleagues compared physical interactions among different cell-types and demonstrated that while promoter-enhancer level chromatin interactions change considerably, the large-scale structures[44] remain intact. To this end, studies based on FISH, 3C, and Hi-C have investigated chromatin interaction patterns at individual genomic loci and observed that a vast majority of these chromatin interactions are constrained within hundreds of kilobases to few megabases and are generally intra-TAD[20,44]. For example, a 1Mb intra-TAD chromatin interaction loop originating from a distal enhancer is known to regulate the Sonic Hedgehog gene (SHH), an essential gene for proper limb development[48]. More recently, Sanyal and colleagues studied the structural patterns of promoters in the ENCODE regions and showed that genes can interact with multiple distal elements, and distal elements loop to multiple genes[20]. This suggests that chromatin interactions at the sub-TAD level can not only be dynamic among different cells, but can be of complex 3D structural pattern in itself enabling combinatorial interactions among genes and regulatory elements.

In this section, I have presented a hierarchical view of genome organization. In particular, CTs form the lower level resolution, while TADs form mid-level, and individual chromatin interactions among genomic loci form the

high-resolution structural patterns. Each of these layers of 3D genome seems to play a critical role in controlling transcription. While we see a clear genome structure and function correlation, understanding genome sequence in this context can allow better understanding of genome function. For example, understanding enrichment of DNA binding protein CTCF at TAD boundaries[36] can explain formation of TADs and potentially their function. Similarly, in order to delineate how disease-associated alleles regulate target genes, an understanding of interplay among genome sequence and the structure is important. Such a combined model can also help in revealing the complex combinatorial patterns of transcriptional activity.

3D genome structure can reveal haplotype patterns

Recent advances in genome-editing tools such as CRISPR have enabled systematic perturbation of genetic sequences, offering an elegant way to assess the genetic background of genome structure and function[49,50]. However, genome-editing tools are currently low-throughput and are laborious to perform. Alternatively, as humans inherit two copies, or haplotypes, of genetic content, sequence differences among the homologous chromosomes can be exploited as natural genetic perturbations, allowing us perform analyses on genome structure and function in high-throughput. Nevertheless, as current genomic DNA sequencing technologies utilize mixtures of maternal and paternal chromosomes that are fragmented prior to sequencing, our ability to distinguish the two haplotypes is extremely limited. In particular, computational approaches can be

used to reconstruct and assemble haplotypes, but they can recover haplotype blocks that are only tens to hundreds of kilobases long[51-55]. In complex genomes such as humans, genetic or epigenetic changes at regulatory sequences can regulate genes much further away, emphasizing the need for obtaining chromosome-span haplotypes[20]. While several experimental approaches[56-58] can generate complete haplotypes, they require equipment not generally available in most research or clinical laboratories or are not applicable to general population[59].

We developed a strategy called HaploSeq[22], to reconstruct chromosome-span haplotypes. Previously, proximity-ligation approaches such as Hi-C[35], 5C[36], 4C[34], and 3C[32], were used solely for investigating spatial relationship between genomic sequences. HaploSeq repurposes Hi-C towards achieving whole genome haplotyping. A fundamental aspect of 3D genome that allows capturing haplotypes is the presence of chromosome territories, where even the homologous chromosomes seem to occupy distinct spatial localization[23]. In particular, as Hi-C captures the spatial configuration of genomic loci, it also preferentially links DNA variants in the same haplotype and therefore preserves haplotype information. We employed computational approaches based on Max-cut graph algorithm[60] to eliminate inter-haplotype sequencing error patterns, and predicted accurate haplotype structures for >80% of alleles in both mouse and human cells. As a result of generating complete haplotypes, Hi-C not only can reveal spatial interactions among genes and regulatory elements, but it can also inform which homologous copy these elements belong to.

While we used Hi-C to deconvolute haplotype patters, other groups have performed de novo assembly[61,62] using these datasets. In addition, studies have shown the utility of 4C towards typing structural variants such as large insertions, inversions and translocations[63]. With myriad of utilities towards analyses of genome structure and sequence, C-based technologies such as 4C, 5C and Hi-C will perhaps be applicable to a wide range of genomic studies in the future.

Gene regulation in an allele-specific context

Previous studies have correlated changes in chromatin structure to gene expression across various cell-types or specific experimental conditions[20, 36, 41, 43, 44, 46]. Similarly, changes in genetic sequence and epigenetic activity have been studied in the context of gene regulation[12, 64-70]. However, studies that integrate many types of information such as epigenetics, haplotypes and chromatin structure have largely been absent, owing to the difficulty in obtaining these datasets. Such integrative studies can substantially advance our knowledge of gene regulatory mechanisms in human cells.

As projects from our lab have demonstrated the utility of Hi-C in delineating chromatin interactions between regulatory elements[36,44] and reconstructing haplotypes[22], we performed Hi-C across embryonic stem cells (ES) and a panel of ES-derived differentiated cells from the H1 human cell line. This system has also been extensively profiled by the Roadmap epigenome project for several epigenetic marks, using which we and other groups have

annotated chromatin states such as enhancers, promoters, insulators, and gene activity across these cell-types[14]. By integrating chromatin states, 3D genome, and haplotype information to this system, we anticipate to explore allelic patterns of gene regulation.

Using HaploSeq[22], we phased 93.5% of alleles to chromosome-spanning haplotypes. With the majority of alleles phased, our study is applicable to genome-wide analyses of allele specific gene expression and underlying chromatin state patterns through cellular differentiation. The haplotype phase resolved genome revealed widespread allele specific gene expression patters, which appears to be strongly correlated with allelic chromatin states of promoters or distal acting enhancers. By adding 3D structure information, we observed that spatially proximal allelic enhancers are strongly correlated to target gene expression. While we cannot determine if the allelic activities are due to genetic or epigenetic factors, our study demonstrates the combinatorial functional aspects of genetic sequence and structure towards gene regulation.

Conclusion

To understand how a cell interprets it genetic content, we must first obtain genetic sequence and annotate the different functional elements. Recent collaborative projects such as ENCODE[9, 16, 17] and Roadmap Epigenome[10-12, 14, 15] have used genome-sequencing tools to profile transcription factor binding, gene expression, chromatin accessibility and epigenetic marks. These datasets have been used to comprehensively map functional elements such as

enhancers, and promoters and have subsequently been used to study transcription[10]. However as humans inherit two copies of genetic content, any genetic or epigenetic difference between the two haplotypes is ignored, limiting our understanding of gene regulation. Further, by exploiting these differences and by adding knowledge of higher-order chromatin interactions to link various regulatory elements and target genes, we can explore novel insights on the landscape of allelic gene regulation patterns. For example, our study on haplotype-resolved H1 genome revealed aspects of distal gene regulation. In particular, compound heterozygosity of distal non-coding alleles can impact transcription and this emphasizes the need for long-range haplotypes as well as 3D genome information. By expanding such integrative analyses to many individuals across different conditions such as disease states or tissue types, we can potentially generate predictive models of the genetic basis of human development and disease.

References

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla,

A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J. & International Human Genome Sequencing, C. Initial sequencing and analysis of the human genome. Nature 409, 860-921 (2001).

2. Pushkarev, D., Neff, N.F. & Quake, S.R. Single-molecule sequencing of an individual human genome. Nature biotechnology 27, 847-850 (2009).

3. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M.,

Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. & Zhu, X. The sequence of the human genome. Science 291, 1304-1351 (2001).

4. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X.Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A. & Rothberg, J.M. The complete genome of an individual by massively parallel DNA sequencing. Nature 452, 872-876 (2008).

5. Ben-Porath, I., Thomson, M.W., Carey, V.J., Ge, R., Bell, G.W., Regev, A. & Weinberg, R.A. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nature genetics 40, 499-507 (2008).

6. Graf, T. & Enver, T. Forcing cells to change lineages. Nature 462, 587-594 (2009).

7. Schnabel, M., Marlovits, S., Eckhoff, G., Fichtel, I., Gotzen, L., Vecsei, V. & Schlegel, J. Dedifferentiation-associated changes in morphology and gene expression in primary human articular chondrocytes in cell culture. Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society 10, 62-70 (2002).

8. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G.H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Jonasdottir, A., Styrkarsdottir, U., Gretarsdottir, S., Magnusson, K.P., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H.G., Stefansson, T., Leifsson, B.G., Thorsteinsdottir, U., Lamb, J.R., Gulcher, J.R., Reitman, M.L., Kong, A., Schadt, E.E. & Stefansson, K. Genetics of gene expression and its effect on disease. Nature 452, 423-428 (2008).

9. Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A.P., Cayting, P., Charos, A., Chen, D.Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E.C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T.E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K.Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P.J., Myers, R.M., Weissman, S.M. & Snyder, M. Architecture of the human regulatory network derived from ENCODE data. Nature 489, 91-100 (2012).

10. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. & Bernstein, B.E. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43-49 (2011).

11. Gifford, C.A., Ziller, M.J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A.K., Kelley, D.R., Shishkin, A.A., Issner, R., Zhang, X., Coyne, M., Fostel, J.L., Holmes, L., Meldrim, J., Guttman, M., Epstein, C., Park, H., Kohlbacher, O., Rinn, J., Gnirke, A., Lander, E.S., Bernstein, B.E. & Meissner, A. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell 153, 1149-1163 (2013).

12. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J.,

Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B. & Ecker, J.R. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315-322 (2009).

13. Maunakea, A.K., Chepelev, I. & Zhao, K. Epigenome mapping in normal and disease States. Circulation research 107, 327-339 (2010).

14. Xie, W., Schultz, M.D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J.W., Tian, S., Hawkins, R.D., Leung, D., Yang, H., Wang, T., Lee, A.Y., Swanson, S.A., Zhang, J., Zhu, Y., Kim, A., Nery, J.R., Urich, M.A., Kuan, S., Yen, C.A., Klugman, S., Yu, P., Suknuntha, K., Propson, N.E., Chen, H., Edsall, L.E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W.Y., Chi, N.C., Antosiewicz-Bourget, J.E., Slukvin, I., Stewart, R., Zhang, M.Q., Wang, W., Thomson, J.A., Ecker, J.R. & Ren, B. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell 153, 1134-1148 (2013).

15. Zhu, J., Adli, M., Zou, J.Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P.L., Bennett, D.A., Houmard, J.A., Muoio, D.M., Onder, T.T., Camahort, R., Cowan, C.A., Meissner, A., Epstein, C.B., Shoresh, N. & Bernstein, B.E. Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell 152, 642-654 (2013).

16. Consortium, E.P., Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C. & Snyder, M. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74 (2012).

17. Consortium, E.P., Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., Koch, C.M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J.A., Andrews, R.M., Flicek, P., Boyle, P.J., Cao, H., Carter, N.P., Clelland, G.K., Davis, S., Day, N., Dhami, P., Dillon, S.C., Dorschner, M.O., Fiegler, H., Giresi, P.G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K.D., Johnson, B.E., Johnson, E.M., Frum, T.T., Rosenzweig, E.R., Karnani, N., Lee, K., Lefebvre, G.C., Navas, P.A., Neri, F., Parker, S.C., Sabo, P.J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F.S., Dekker, J., Lieb, J.D., Tullius, T.D., Crawford, G.E., Sunyaev, S., Noble, W.S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I.L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H.A., Sekinger, E.A., Lagarde, J., Abril, J.F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J.S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M.C., Thomas, D.J., Weirauch, M.T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K.G., Sung, W.K., Ooi,

H.S., Chiu, K.P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M.L., Valencia, A., Choo, S.W., Choo, C.Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T.G., Brown, J.B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C.N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J.S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R.M., Rogers, J., Stadler, P.F., Lowe, T.M., Wei, C.L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S.E., Fu, Y., Green, E.D., Karaoz, U., Siepel, A., Taylor, J., Liefer, L.A., Wetterstrand, K.A., Good, P.J., Feingold, E.A., Guyer, M.S., Cooper, G.M., Asimenos, G., Dewey, C.N., Hou, M., Nikolaev, S., Montoya-Burgos, J.I., Loytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N.R., Holmes, I., Mullikin, J.C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W.J., Stone, E.A., Program, N.C.S., Baylor College of Medicine Human Genome Sequencing, C., Washington University Genome Sequencing, C., Broad, I., Children's Hospital Oakland Research, I., Batzoglou, S., Goldman, N., Hardison, R.C., Haussler, D., Miller, W., Sidow, A., Trinklein, N.D., Zhang, Z.D., Barrera, L., Stuart, R., King, D.C., Ameur, A., Enroth, S., Bieda, M.C., Kim, J., Bhinge, A.A., Jiang, N., Liu, J., Yao, F., Vega, V.B., Lee, C.W., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M.J., Inman, D., Singer, M.A., Richmond, T.A., Munn, K.J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J.C., Couttet, P., Bruce, A.W., Dovey, O.M., Ellis, P.D., Langford, C.F., Nix, D.A., Euskirchen, G., Hartman, S., Urban, A.E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T.H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C.K., Rosenfeld, M.G., Aldred, S.F., Cooper, S.J., Halees, A., Lin, J.M., Shulha, H.P., Zhang, X., Xu, M., Haidar, J.N., Yu, Y., Ruan, Y., Iyer, V.R., Green, R.D., Wadelius, C., Farnham, P.J., Ren, B., Harte, R.A., Hinrichs, A.S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A.S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R.M., Karolchik, D., Armengol, L., Bird, C.P., de Bakker, P.I., Kern, A.D., Lopez-Bigas, N., Martin, J.D., Stranger, B.E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrimsdottir, I.B., Huppert, J., Zody, M.C., Abecasis, G.R., Estivill, X., Bouffard, G.G., Guan, X., Hansen, N.F., Idol, J.R., Maduro, V.V., Maskeri, B., McDowell, J.C., Park, M., Thomas, P.J., Young, A.C., Blakesley, R.W., Muzny, D.M., Sodergren, E., Wheeler, D.A., Worley, K.C., Jiang, H., Weinstock, G.M., Gibbs, R.A., Graves, T., Fulton, R., Mardis, E.R., Wilson, R.K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D.B., Chang, J.L., Lindblad-Toh, K., Lander, E.S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B. & de Jong, P.J. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799-816 (2007).

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

19. Kosak, S.T. & Groudine, M. Form follows function: The genomic organization of cellular differentiation. Genes & development 18, 1371-1384 (2004).

20. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. Nature 489, 109-113 (2012).

21. Solovei, I., Cavallo, A., Schermelleh, L., Jaunin, F., Scasselati, C., Cmarko, D., Cremer, C., Fakan, S. & Cremer, T. Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). Experimental cell research 276, 10-23 (2002).

22. Selvaraj, S., J, R.D., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nature biotechnology 31, 1111-1118 (2013).

23. Cremer, T., Cremer, M., Dietzel, S., Muller, S., Solovei, I. & Fakan, S. Chromosome territories--a functional nuclear landscape. Current opinion in cell biology 18, 307-316 (2006).

24. Lichter, P., Ledbetter, S.A., Ledbetter, D.H. & Ward, D.C. Fluorescence in situ hybridization with Alu and L1 polymerase chain reaction probes for rapid characterization of human chromosomes in hybrid cell lines. Proceedings of the National Academy of Sciences of the United States of America 87, 6634-6638 (1990).

25. Pinkel, D., Landegent, J., Collins, C., Fuscoe, J., Segraves, R., Lucas, J. & Gray, J. Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. Proceedings of the National Academy of Sciences of the United States of America 85, 9138-9142 (1988).

26. Janicki, S.M., Tsukamoto, T., Salghetti, S.E., Tansey, W.P., Sachidanandam, R., Prasanth, K.V., Ried, T., Shav-Tal, Y., Bertrand, E., Singer, R.H. & Spector, D.L. From silencing to gene expression: real-time analysis in single cells. Cell 116, 683-698 (2004).

27. Tsukamoto, T., Hashiguchi, N., Janicki, S.M., Tumbar, T., Belmont, A.S. & Spector, D.L. Visualization of gene activity in living cells. Nature cell biology 2, 871-878 (2000).

28. Branco, M.R. & Pombo, A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS biology 4, e138 (2006).

29. Pickersgill, H., Kalverda, B., de Wit, E., Talhout, W., Fornerod, M. & van Steensel, B. Characterization of the Drosophila melanogaster genome at the nuclear lamina. Nature genetics 38, 1005-1014 (2006).

30. van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nature biotechnology 18, 424-428 (2000).

31. McDonald, O.G., Wu, H., Timp, W., Doi, A. & Feinberg, A.P. Genome-scale epigenetic reprogramming during epithelial-to-mesenchymal transition. Nature structural & molecular biology 18, 867-874 (2011).

32. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. Science 295, 1306-1311 (2002).

33. Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., Green, R.D. & Dekker, J. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome research 16, 1299-1309 (2006).

34. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. & de Laat, W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nature genetics 38, 1348-1354 (2006).

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

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

37. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature genetics 43, 1059-1065 (2011).

38. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. & Cavalli, G. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 148, 458-472 (2012).

39. Munkel, C., Eils, R., Dietzel, S., Zink, D., Mehring, C., Wedemann, G., Cremer, T. & Langowski, J. Compartmentalization of interphase chromosomes

observed in simulation and experiment. Journal of molecular biology 285, 1053-1065 (1999).

40. Yokota, H., van den Engh, G., Hearst, J.E., Sachs, R.K. & Trask, B.J. Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus. The Journal of cell biology 130, 1239-1249 (1995).

41. Chambeyron, S. & Bickmore, W.A. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. Genes & development 18, 1119-1130 (2004).

42. Clowney, E.J., LeGros, M.A., Mosley, C.P., Clowney, F.G., Markenskoff-Papadimitriou, E.C., Myllys, M., Barnea, G., Larabell, C.A. & Lomvardas, S. Nuclear aggregation of olfactory receptor genes governs their monogenic expression. Cell 151, 724-737 (2012).

43. Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W. & Duboule, D. The dynamic architecture of Hox gene clusters. Science 334, 222-225 (2011).

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

45. Hu, M., Deng, K., Qin, Z., Dixon, J., Selvaraj, S., Fang, J., Ren, B. & Liu, J.S. Bayesian inference of spatial organizations of chromosomes. PLoS computational biology 9, e1002893 (2013).

46. Bau, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J. & Marti-Renom, M.A. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. Nature structural & molecular biology 18, 107-114 (2011).

47. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A. & Fraser, P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502, 59-64 (2013).

48. Maas, S.A. & Fallon, J.F. Single base pair change in the long-range Sonic hedgehog limb-specific enhancer is a genetic basis for preaxial polydactyly. Developmental dynamics : an official publication of the American Association of Anatomists 232, 345-348 (2005).

49. Jinek, M., East, A., Cheng, A., Lin, S., Ma, E. & Doudna, J. RNA-programmed genome editing in human cells. eLife 2, e00471 (2013).

50. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. & Church, G.M. RNA-guided human genome engineering via Cas9. Science 339, 823-826 (2013).

51. Kitzman, J.O., Mackenzie, A.P., Adey, A., Hiatt, J.B., Patwardhan, R.P., Sudmant, P.H., Ng, S.B., Alkan, C., Qiu, R., Eichler, E.E. & Shendure, J. Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nature biotechnology 29, 59-63 (2011).

52. Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., Haas, J., Robasky, K., Zaranek, A.W., Lee, J.H., Ball, M.P., Peterson, J.E., Perazich, H., Yeung, G., Liu, J., Chen, L., Kennemer, M.I., Pothuraju, K., Konvicka, K., Tsoupko-Sitnikov, M., Pant, K.P., Ebert, J.C., Nilsen, G.B., Baccash, J., Halpern, A.L., Church, G.M. & Drmanac, R. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487, 190-195 (2012).

53. Suk, E.K., McEwen, G.K., Duitama, J., Nowick, K., Schulz, S., Palczewski, S., Schreiber, S., Holloway, D.T., McLaughlin, S., Peckham, H., Lee, C., Huebsch, T. & Hoehe, M.R. A comprehensively molecular haplotype-resolved genome of a European individual. Genome research 21, 1672-1685 (2011).

54. Kaper, F., Swamy, S., Klotzle, B., Munchel, S., Cottrell, J., Bibikova, M., Chuang, H.Y., Kruglyak, S., Ronaghi, M., Eberle, M.A. & Fan, J.B. Whole-genome haplotyping by dilution, amplification, and sequencing. Proceedings of the National Academy of Sciences of the United States of America 110, 5552-5557 (2013).

55. Duitama, J., McEwen, G.K., Huebsch, T., Palczewski, S., Schulz, S., Verstrepen, K., Suk, E.K. & Hoehe, M.R. Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. Nucleic acids research 40, 2041-2053 (2012).

56. Fan, H.C., Wang, J., Potanina, A. & Quake, S.R. Whole-genome molecular haplotyping of single cells. Nature biotechnology 29, 51-57 (2011).

57. Yang, H., Chen, X. & Wong, W.H. Completely phased genome sequencing through chromosome sorting. Proceedings of the National Academy of Sciences of the United States of America 108, 12-17 (2011).

58. Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K. & Song, Q. Direct determination of molecular haplotypes by chromosome microdissection. Nature methods 7, 299-301 (2010).

59. Kirkness, E.F., Grindberg, R.V., Yee-Greenbaum, J., Marshall, C.R., Scherer, S.W., Lasken, R.S. & Venter, J.C. Sequencing of isolated sperm cells for direct haplotyping of a human genome. Genome research 23, 826-832 (2013).

60. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics 24, i153-159 (2008).

61. Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. & Shendure, J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature biotechnology 31, 1119-1125 (2013).

62. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from in vivo DNA interaction frequency. Nature biotechnology 31, 1143-1147 (2013).

63. Simonis, M., Klous, P., Homminga, I., Galjaard, R.J., Rijkers, E.J., Grosveld, F., Meijerink, J.P. & de Laat, W. High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. Nature methods 6, 837-842 (2009).

64. Xie, W., Barr, C.L., Kim, A., Yue, F., Lee, A.Y., Eubanks, J., Dempster, E.L. & Ren, B. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell 148, 816-831 (2012).

65. Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V. & Ren, B. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116-120 (2012).

66. Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., Wysocka, J., Lei, M., Dekker, J., Helms, J.A. & Chang, H.Y. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472, 120-124 (2011).

67. Wen, B., Wu, H., Shinkai, Y., Irizarry, R.A. & Feinberg, A.P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. Nature genetics 41, 246-250 (2009).

68. Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K.T., Jonasdottir, A., Frigge, M.L., Gylfason, A., Olason, P.I., Gudjonsson, S.A., Sverrisson, S., Stacey, S.N., Sigurgeirsson, B., Benediktsdottir, K.R., Sigurdsson, H., Jonsson, T., Benediktsson, R., Olafsson, J.H., Johannsson, O.T., Hreidarsson, A.B., Sigurdsson, G., Consortium, D., Ferguson-Smith, A.C., Gudbjartsson, D.F., Thorsteinsdottir, U. & Stefansson, K. Parental origin of sequence variants associated with complex diseases. Nature 462, 868-874 (2009).

69. Krueger, C., King, M.R., Krueger, F., Branco, M.R., Osborne, C.S., Niakan, K.K., Higgins, M.J. & Reik, W. Pairing of homologous regions in the mouse genome is associated with transcription but not imprinting status. PloS one 7, e38983 (2012).

70. McDaniell, R., Lee, B.K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., Keefe, D., Collins, F.S., Willard, H.F., Lieb, J.D., Furey, T.S., Crawford, G.E., Iyer, V.R. & Birney, E. Heritable individual-specific and allele-specific chromatin signatures in humans. Science 328, 235-239 (2010).

Chapter 2: Mammalian genomes are organized into topological domains

Abstract

The 3D structure of the genome occupies distinct chromosome territories, but how chromatin folds within these territories is poorly understood. A common feature of several theoretical models suggests a domain-like organization of chromatin folding, but the exact size and boundaries of these domains have not been well defined. By using the Hi-C protocol, we profiled the genome-wide chromatin interactions in human and mouse embryonic stem cells, and a panel of terminally differentiated cell types. Our initial analyses of the data revealed the presence of highly self-interacting and spatially isolated regions, which we termed as topological domains (TADs). I developed a hidden-markov model based algorithm to show that the mammalian chromosomes are segmented into megabase-sized TADs. We also found that the TADs are pervasive throughout the genome, stable across different cell-types, and conserved between mouse and human. In addition, topological domain boundaries appear to mark the transition between active and inactive regions of the genome, as observed by enrichment of H3K9me3 and its relatedness to A/B compartments and Lamina-associated domains. Further, I have developed statistical methods to correlate cell-type specific chromatin interactions to cell-type specific gene expression, illustrating coupled activity between genome structure and function.

Introduction

Nearly all cells in a mammalian organism carry the same genetic content and yet functional diversity exists among various cell or tissue types[1,2]. Cells achieve this diversity by regulating different subset of genes, which is facilitated and accompanied by coordinated changes in 3D genome or chromatin structure[2-4]. For instance, previous studies have shown that chromatin loop interactions between promoters and distal regulatory elements such as enhancers are critical for gene activation[5-7]. In another instance, Stavros Lomvardas and colleagues used X-Ray tomography to show that olfactory receptor genes from different chromosomes assemble in a few heterochromatic loci, demonstrating co-regulation of genes across multiple chromosomes[8]. Understanding the higher order chromatin structure is therefore essential in comprehending how genes are regulated, which in turn can further our knowledge in cell development and disease.

In eukaryotic cells, the higher order structure of the genome is organized at multiple levels[3,9]. Specifically, it has been suggested that chromosomes occupy distinct regions in the interphase nucleus called chromosome territories (CTs), but our view of the chromosome folding within these CTs are coarse and incomplete. Several models have been suggested to describe these structures, including random-walk/giant loop model[3,10], chromatin rosette/short loop model[4] and more recently fractal globule conformation[11]. A converging aspect of these models is recurring loops or domains of chromatin organization, however, the location, size, and properties of these domains have not been well studied.

Previous groups have linked the domain-like organization of genome structure to transcription for a few genomic loci[8,12-14]. A well-known example is the Fluorescent in situ hybridization (FISH) based study that demonstrated Hoxb domain condensation inside and outside of CTs, correlating well with gene expression[14]. In another example, Bau and colleagues used chromosome conformation capture based 5C technique to show the functional impact of structural changes between human GM12878 and K562 cells at α-globin domain[13]. As such techniques such as FISH[15] and chromosome conformation capture[16] are low-throughput and does not enable genome-wide understanding of the relationship between the higher order chromatin structure and genome function.

Recently, Job Dekker and colleagues have introduced Hi-C[11], as a genome-wide extension of chromosome conformation capture (3C). Hi-C relies on proximity ligation followed by PCR and high-throughput sequencing to assess the spatial relationship between all pairs of genomic loci in vivo. The spatial proximity is inversely proportional to the contact frequencies (# of reads) between two fragments. In this study, we performed Hi-C in multiple human and mouse cells to define the location of domains and to characterize the 3D structure of genome in relation to its function. I used a hidden-markov based algorithm and found that the mammalian genome is organized in to more than a thousand megabase-sized topological domains or TADs. I also investigated how these domain structures change conformation through differentiation and correlated these to changes in gene regulation. In addition, TADs appear to be stable

across cell types, and are highly conserved across species, suggesting that TADs are an inherent property of mammalian genomes.

Results

Hi-C analyses in human and mouse cells

Our lab performed at least two replicates of Hi-C, each in human and mouse embryonic stem cells (ES), and terminally differentiated human IMR90 as well as mouse cortex cells[17]. Together, we analyzed over 1.7 billion paired-end reads of Hi-C data. As a first step, we validated our Hi-C data with previously published chromatin interaction datasets. In particular, both replicates of our IMR90 Hi-C data showed high degree of similarity when compared to 5C dataset from lung fibroblasts (Fig. 2-1)[18]. Further, our mouse ES Hi-C data recovered previously described cell-type specific interaction at the Phc1 locus[19] (Fig. 2-2).

As Hi-C measures spatial proximity among all pairs of loci, significant differences in genomic properties among various loci can potentially generate systemic variability in the data. Therefore, I implemented the recently published probabilistic method to normalize Hi-C data[20]. In brief, genomic loci were binned based on properties such as GC content, mappability, and restriction fragment length, and together these were non-parametrically modeled to enrich chromatin interaction signals over noise (Methods). While Hi-C interaction counts clearly depend on the frequency of restriction enzyme cut sites prior to normalization, the biases have been largely eliminated after normalization (Fig. 2-3). In addition, the correlation between Hi-C Nco1 mouse ES data and previously published

FISH dataset[12] phenomenally increases after normalization (Fig. 2-4), demonstrating that the normalized Hi-C data can accurately reproduce the expected spatial distance from an independent method. These results demonstrate that our Hi-C data across multiple replicates among various cell-types are of high quality.

Identification of Topological domains

One striking feature of the Hi-C data when visualized as a two-dimensional matrix of 40-kb genomic bins is the prevalence of genomic regions displaying high frequencies of local interactions (Fig. 2-5a), seen as “triangles” on the matrix. We hypothesized that these local regions of high frequency interactions represent higher order interacting topological domains, or “TADs”. In addition, narrow segments bound topological domains where the chromatin interactions appear to end abruptly (Fig. 2-5a) and we believed that these abrupt transitions might represent boundary regions that separate topological domains. Furthermore, bins flanking boundaries are biased towards interacting either upstream or downstream depending upon whether they are upstream or downstream to the boundary. We hence hypothesized that there are genomic regions that are specifically biased in upstream vs. downstream and vice versa, and that by detecting these locally biased regions, we would be able to objectively identify the location of topological domains.

We expected each bin to be unbiased (as null hypothesis) and we asked for a quantification of the degree of bias using chi-square statistic for every bin in

a given chromosome. In particular, for every 40-kb bin of the genome, we looked 2-Mb upstream and 2-Mb downstream to estimate chi-square based biases. We labeled the upstream biases as negative and downstream biases as positive (Fig. 2-5c,e). We called the degree of bias as the directionality index (DI) and as described earlier, we notice the directionality index changes abruptly at the boundaries and that the domains appear to contain a cluster of downstream biased bins followed by cluster of upstream biased bins.

As DI quantifies the degree of bias of a given bin, we observe that for most of the bins, the DI accounts values close to 0 and therefore does not clearly pinpoint the bias (Fig 2-5f). As Hi-C is performed in million of cells and that these cells are unsynchronized in their cell-cycle stages, DI can be affected by stochasticity. Hence, we were in need of a system that considers the DI as observations, models them to account for variation and noise, and predicts whether a region could be upstream biased, downstream biased, or not biased. Since every bin in the genome has an unknown state and that the previous bin influences current bin (due to the clustering property of DI), I developed a hidden markov model (HMM) based algorithm that estimates the “true” directionality bias of every bin in the genome given the DI observations (Fig. 2-6). Specifically, the HMM assumes that the DI observations are following a mixture of Gaussians and then predicts the states as “Upstream Bias”, “Downstream Bias” or “No Bias”.

For the HMM algorithm, I concatenated the DI’s across a given chromosome and assuming it is a vector of size n, where n= size of chromosome/bin size. For instance, describing the observed DI’s as Y’s

$[Y_1,Y_2..Y_n]$, the hidden true directionality biases as Q's $[Q_1,Q_2..Q_n]$ and the mixtures as M's $[M_1,M_2..M_n]$. The probability $P(Y_t|Q_t = i,M_t = m)$ is represented using a mixture of Gaussians for each state i. The Conditional probability distribution [CPDs] of Y_t and M_t nodes are defined as,

$P(Y_t = y_t|Q_t = i,M_t = m) = N(y_t;\mu_{i,m},\Sigma_{i,m})$

$P(M_t =m|Q_t =i) = C(i,m)$, where C is the mixture weights for each state i.

I used Baum-Welch algorithm based on Expected Maximization principles to compute maximum likelihood estimate and the parameter estimates of transition and emission (characterized by mean, covariance and weights). The posterior marginals were then estimated using the Forward-backward algorithm. I predicted the HMM states by allowing 1 to 20 mixtures. I chose the mixture with best goodness of fit using the AIC criterion, $AIC = 2k – 2\ln(L)$, k is the number of parameters in the model and L being the maximum likelihood estimate. In summary, for each chromosome we fit the HMM model with best suiting mixtures of varying sizes from 1 to 20. More recently, I have modified the algorithm to fit the HMM model for the DIs from all the chromosomes together, to utilize the entirely of data for estimating best suiting mixture, M. In addition, instead of choosing M with the lowest AIC as the best goodness of fit, a model with at most 10% loss of AIC seemed to generate consistent results with lesser parameters. The updated version of domain calling algorithm is available to download (Methods).

As a post-processing step, I estimated the median posterior probability of a region, defined as a stretch of same state, and considered only in regions

having a median posterior marginal probabilities ≥ 0.99 or a region that is at least 80-kb (2 bins) long. Domains and boundaries are then inferred from the results of the HMM state calls throughout the genome (Fig. 2-5d). A domain is initiated at the beginning of a single downstream biased HMM state. The domain is continuous throughout any consecutive downstream biased states. The domain will then end when the last in a series of upstream biased states are reached, with the domain ending at the end of the last HMM upstream biased state (Fig. 2-5b). We term the regions in between the topological domains as either “boundaries” or “unorganized chromatin.” We defined unorganized chromatin to be these regions that are > 400kb, and the boundaries to be less than 400kb.

HMM based domain boundary calls are robust

The domain boundaries defined by HMM (Fig. 2-6) were highly reproducible between replicates (Fig. 2-7). Therefore, I combined the data from the HindIII replicates and identified 2,200 topological domains in mouse ES cells with a median size of 880kb that occupy 91% of the genome (Fig. 2-8a). In addition, the median boundary size were ~0 base-pairs and that 76.3% of the boundaries were less than 50 kilobases, indicating that the domain boundary identification by the HMM model were precise (Fig. 2-8b). The median size of unorganized chromatin were ~560 kilobases (Fig. 2-8c). On the same lines, I identified over a 1000 domains each in mouse cortex, human ES and human IMR90 cells using combined datasets from replicates.

As another measurement of robustness in domain identification, we checked the frequency of intra-domain interactions and as expected these were higher than inter-domain interactions (Fig. 2-9a). Similarly, FISH probes in the same topological domain (Fig. 2-9b) are closer in nuclear space than probes in different topological domains (Fig. 2-9c), despite similar genomic distances between probe pairs[12] (Fig. 2-9d-e). These findings are best explained by a model of the organization of genomic DNA into spatial modules (TADs) linked by short chromatin segments, which we define as boundaries.

TADs are largely invariant among cell-types

As the topological domain boundaries identified by HMM are reproducible among replicates, I extended this analysis to compare the boundaries among cell-types in both humans and mouse. I observed a high degree of consistency in the boundary regions identified between mouse ES and cortex (Fig. 2-10a) as well as between human ES and IMR90 (Fig. 2-10b). In addition, at the boundaries called in only one cell type, we noticed that trend of upstream and downstream bias in the directionality index is still readily apparent and highly reproducible between replicates (Fig. 2-10c-d). Currently, we cannot determine if the differences in domain calls between cell types is due to noise in the data or to biological phenomena, such as a change in the strength of the boundary region between cell types[21]. Regardless, most of the domains identified are stable across cell-types. Lastly, a very small fraction of the boundaries show clear

differences between cell-types, but it is unclear how this difference in boundary structure imparts changes in genome function (Fig. 2-11a-b, Methods).

Higher-order conformations of TADs can vary among cell-types

While topological domains are largely invariant among cell-types, their conformation or shape might change causing cell-type specific gene regulation patterns[6,7,13,14]. For instance, cell-type specific interactions can lead to different domain conformations and consequently cell-type specific expression in Phc1 gene[19] (Fig. 2-2), while the domain size and locations are consistent. To identify this phenomenon in a genome-wide fashion, I used a binomial distribution to find dynamic interactions between two cell-types. In particular, I combined data from two replicates of mouse ES and cortex and then used binomial distribution for each possible interaction (20-kb bins) in the genome up to a distance of 5 mega-bases.

Mathematically, $n_d = I_{mESC} + I_{mcortex}$, where n = total trials at a distance d and Expectation $p_{mESC,d} = (\Sigma\ I_{mESC})/n$ and $p_{mcortex,d} = 1\text{-}p_{mESC,d}$. As the spatial proximity between two bins depends on the distance between two bins, I chose to fit a binomial distribution for every distance d, where d varies from 20-kb to 5-Mb. Based on the expectations, I calculated deviations in the ratio of the number of interactions in mouse ES cells ($I_{ij\text{-}mESC}$) to the number of interactions in cortex ($I_{ij\text{-}cortex}$) to obtain statistically significant dynamic interactions. We then randomly permuted the replicates (ES-rep1+Cortex-rep1 Vs ES-rep2+Cortex-rep2) and

(ES-rep1+Cortex-rep2 Vs ES-rep2+Cortex-rep1) to estimate a false discovery rate (FDR).

I identified 9,888 dynamic interacting regions in the mouse genome based on 20-kb binning using a binomial test at an FDR of 1%. As expected, the dynamic interactions are enriched for differentially expressed genes (Fig. 2-12a). In addition, ~20% of the genes that are differentially expressed are a part of dynamic interactions (Fig. 2-12b). This is an underestimate given that dynamic interactions are 20-kb bin sizes and those interactions that are less than this resolution will be missed. As ~96% of dynamic interactions are intra-domain (Fig. 2-12c), it appears as though chromatin interactions are constrained within domains by acting as functional modules of genome structure. In addition, it also suggests that while topological domains size and location are consistent, their conformation and shape might vary leading to dynamic gene regulatory patterns driving cell development and disease.

TADs are evolutionarily conserved

Next, we studied the evolutionary conservation of domains across mouse and humans. To address this, I compared the domain boundaries between mouse ES cells and human ES cells using the UCSC liftover tool[22]. Indeed, majority of boundaries appear to be shared across evolution (53.8% of human boundaries are boundaries in mouse and 75.9% of mouse boundaries are boundaries in humans, compared to 21.0% and 29.0% at random, P value = 2.2x10-16, Fisher's exact test; Fig. 2-13a). The random boundaries were

determined by constraining on the distribution of boundary lengths and distribution of chromosomal occurrence. The syntenic regions in mouse and human in particular share a high degree of similarity in their higher order chromatin structure (Fig. 2-13b). This suggests that beyond conservation of sequence elements across evolution, structural features might also be conserved and thus reiterating its likely role in genome function.

Insulator/barrier elements mark TAD boundaries

We observed a strong enrichment of insulator binding element CTCF at the boundary regions of topological domains (Fig. 2-14a). Specifically, >85% of boundaries in mouse ES cells contained CTCF binding site (Fig. 2-14b), re-iterating that boundaries share this property of classical insulator element[23, 24]. In addition, a classical insulator element pre-marks the sites known to stop the spread of heterochromatin. Consequently, we examined the distribution of H3K9me3 in humans at the shared topological domain boundary sites among ES and IMR90[25,26]. Indeed, we observe a clear segregation of H3K9me3 mark at the boundary, predominantly in the differentiated cell type of IMR90 (Fig. 2-14c). Specifically as we analyzed shared boundaries, it seems as though while the boundaries are constant, heterochromatin marks are rewritten in differentiated cell-types (Fig. 2-14f).

Previous studies have reported other means of genome compartmentalization, such as A and B compartments[11] and Lamina-associated domains (LAD)[27,28]. We compared our topological domain definitions to these

strictures and observed that the topological domain boundaries mark the transition of A and B compartments, as well as LAD and non-LADs (Fig. 2-14d-e). Taken together, the above observations strongly suggest that the topological domain boundaries correlate with regions of the genome displaying insulator activity and marks transitions between active and inactive regions of the genome, thus revealing a potential link between genome structure and transcription.

Discussion

In this study, we show that the mammalian chromosomes are segmented into megabase-sized topological domains. Using the HMM based algorithm, we have now been able to determine exact genomic locations and size of these topological domains and boundaries to an unprecedented precision. Such spatial organization appears to be a general property of the genome: it is pervasive throughout the genome, stable across different cell types and are highly conserved between mouse and humans.

We have investigated functional relationship between the topological domains and genome structure in several ways. For one, while the domain location and size are consistent across different cell-types, their conformation seems to change through the presence of dynamic interactions that can in turn allow for cell-type specific gene regulation patterns. Second, as ~96% of dynamic interactions are intra-domain, the topological domains appear to act as functional regulatory modules that restrict chromatin interactions. Third, boundaries of topological domains are associated with the CTCF, suggesting that the

topological domains correspond to insulator or barrier elements of the genome. Fourth, topological domains appear to mark the transition between active and inactive regions of the genome by stopping the spread of heterochromatin as well as by marking A/B[11] and LAD transitions[27,28].

While we and other have observed topological domains in drosophila[29], E. coli[30], mouse[31] and human (our study), and have investigated functional links between genome structure and function[6,8,14], an obvious next step would be to provide mechanistic details on the genome structure-function relationship. For one, genome editing tools such as CRISPR[32,33] and TALEN[34,35] can be used to delete boundaries and can allow prediction of gene regulation. Second, a higher resolution Hi-C dataset[36] or techniques such as ChIA-PET[37] can allow for studying of individual fragment based functional interactions between promoters and regulatory elements, unlike bin based analysis in our study. This is a critical step in assigning target genes for the majority of disease associated non-coding variants[38]. Altogether, determining mechanistic details of genome structure that allows for building predictive models of gene regulation will be an important step in the future.

Methods

Hi-C data mapping to reference genome

We mapped the paired-end Hi-C data as two independent single end reads using BWA[39] with default parameters. We used samtools[40] to consider only

uniquely mapping reads (mapping quality > 10). We removed PCR duplicate reads using Picard (http://picard.sourceforge.net).

Data normalization

I normalized the data as previously described by Yaffe and Tanay[20]. This method works by taking in to account three parameters that impact Hi-C signal – GC content, fragment length and mappability of fragments. Yaffe and Tanay[20] nicely showed that these three parameters interact in a non-linear way. For our implementation of this protocol, I first assigned reads to nearby fragments and then removed all reads that belonged to fragments having mappability score < 0.5. Previously, mappability score is estimated as a fraction of simulated reads that mapped for any given fragment. Next, I binned all reads in 20x20 matrices of fragment length (FL) and GC content each and calculated the probability distributions of the variability of GC and FL to non-parametrically estimate an expectation value for the observed Hi-C signal. Specially, the expectation was calculated for all read-pairs originating from a given 40-kb bin pair. In comparison to Yaffe and Tanay[20], we did not perform linear weight smoothing and BFGS non-linear optimization. Despite this, the normalization method is still effective at removing restriction enzyme bias (Fig. 2-3 and Fig. 2-4).

Resolution of TAD analyses

We chose to work with 40-kb bin sizes for identifying topological domains and 20-kb bin sizes for determining dynamic interactions. Our resolution was determined based on the coverage of Hi-C data generated in this study.

Correlation between Experiments

We calculated the correlation between two experiments as follows: The set of all possible interactions I_{ij} for two experiments A and B were correlated by comparing each point in interaction matrix I_A from experiment A with the same point I_B from experiment B. Because the interaction matrix is highly skewed towards proximal interactions, we restricted the correlation to a maximum distance between points i and j of 50 bins. We use R to calculate the Pearson correlation between the two vectors of all point in I_A and I_B.

Enrichment of factors at boundaries

For determining which boundaries are associated with CTCF, we considered a boundary to be associated CTCF if there were a binding site called by MACS[41] within +/- 20-kb of the boundary. The 20-kb window is chosen because this reflects the inherent uncertainty in the exact position of the domain calls due to 40-kb binning. For H3K9me3 heatmap and LAD analyses, we used k-means clustering to cluster the data within +/- 500-kb of the boundary.

Determining cell-type specific boundaries

We calculated spearman coefficient of the directionality index between two cells. Specifically, if a boundary was called by the HMM in either cell type, we correlate a vector of directionality indexes +/- 10 bins from the center of the boundary between two experiments of interest. For random correlation, we randomly selected 20 bins from each of the two cell types and calculated the spearman correlation between the two vectors. We repeated the randomization 10,000 times to achieve the random distribution of spearman correlation coefficients. Boundaries were called as “cell type specific” if the boundary regions was identified by the HMM domain calling in only one cell and lacked a significant correlation in the directionality index between the two cell types.

Domain calling algorithm

The latest version of the software is available to download from http://bioinformatics-renlab.ucsd.edu/collaborations/sid/domaincall_software.zip

Figures

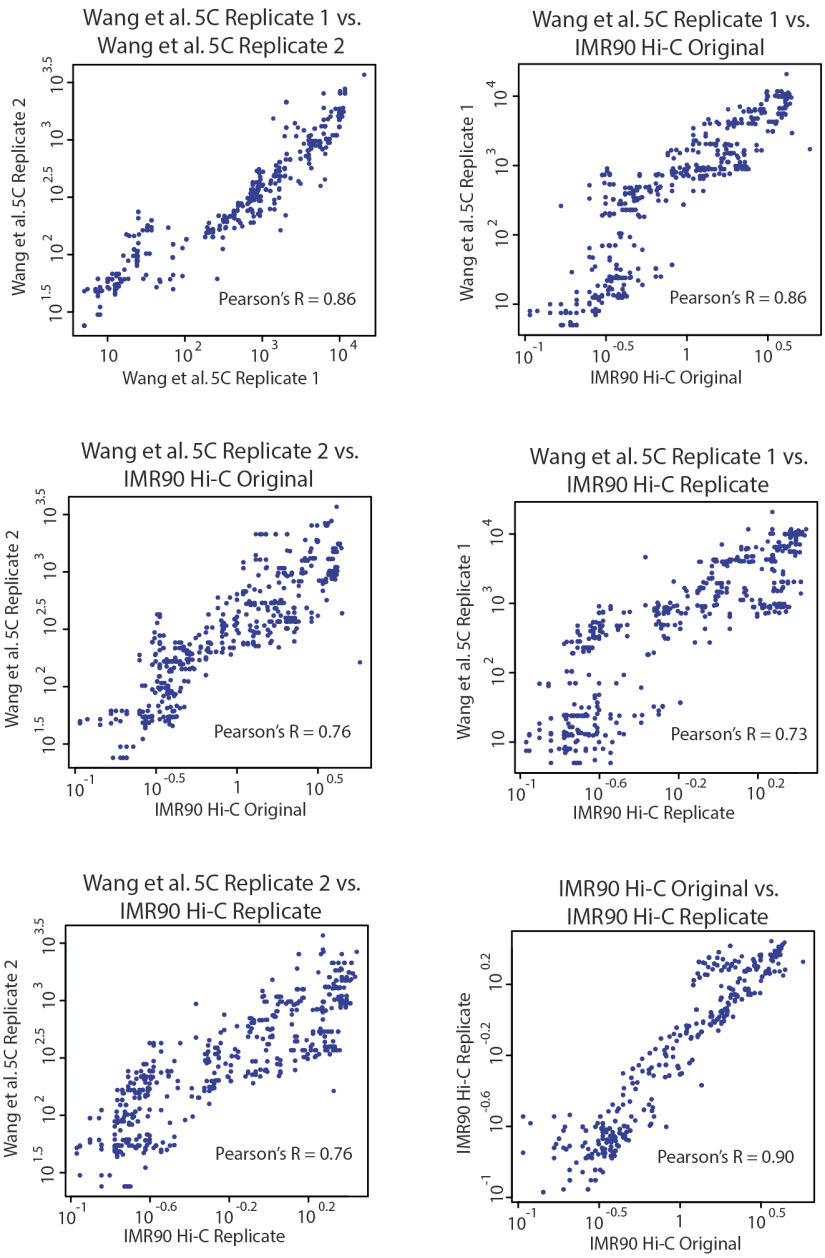

Figure 2-1: Hi-C data correlates well with previously published 5C.

Scatter plots showing the correlations between 5C replicates (ref. 18) and Hi-C data. In all cases, the correlation is > 0.73, demonstrating a high degree of correlation between IMR90 Hi-C data and existing 5C data from a similar cell type.

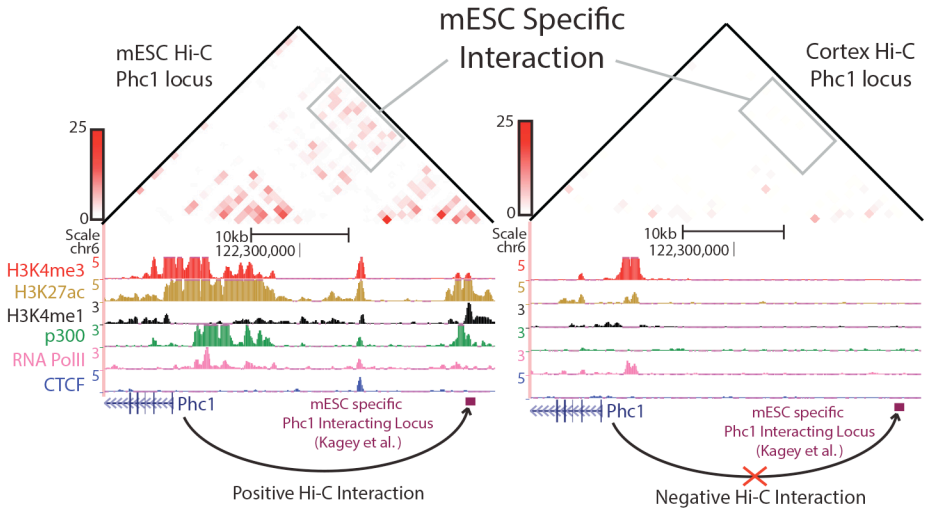

Figure 2-2: Hi-C data recoveres previously described mouse ES specific activity at Phc1 locus.

Our mouse ES and cortex Hi-C data agreed well with previously described (ref. 19) ES specific chromatin interaction and corresponding ES specific Phc1 gene expression activties.

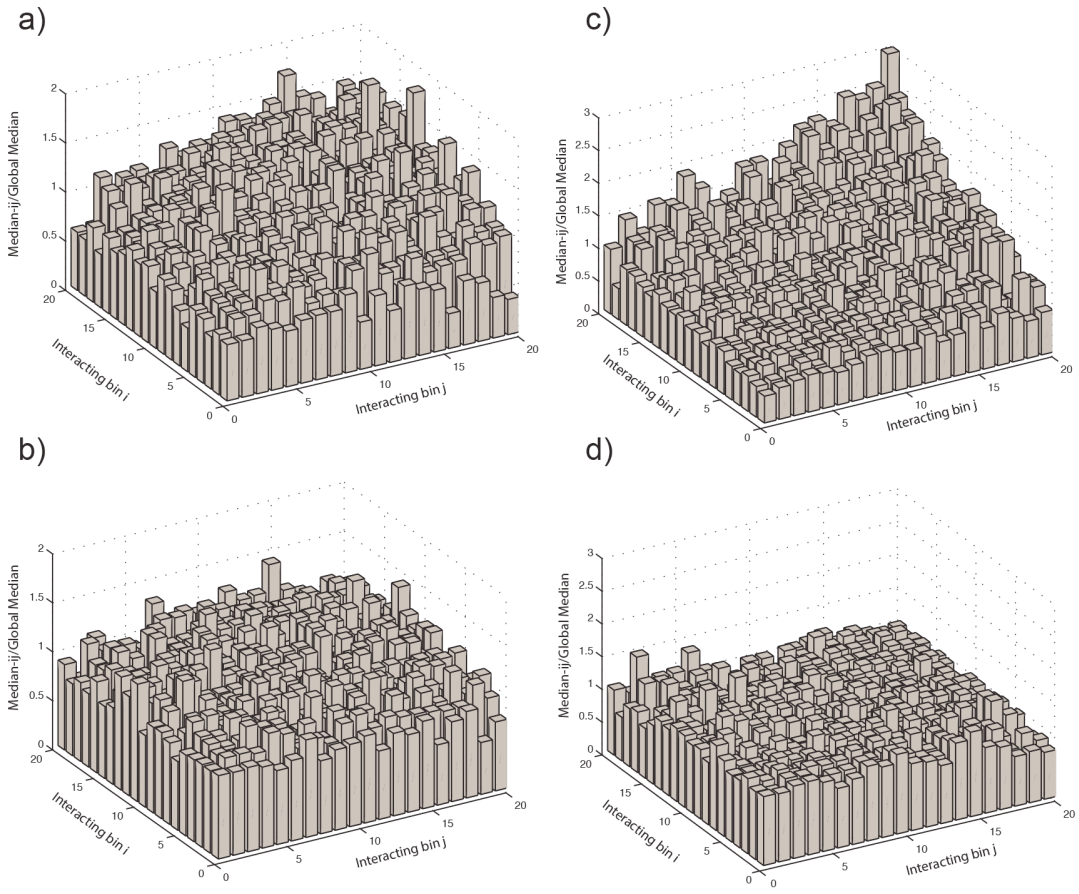

Figure 2-3: Hi-C data biases have been largely reduced after normalization.

Bias plots showing the correlation between restriction enzyme cut site frequency and Hi-C interaction frequency from mouse ES data using a bin size of 250kb at a distance of 1Mb. X and Z axes have bins i and j are grouped into 20 equal sized groups based on increasing restriction enzyme frequency. Y axis shows the median of all interactions Iij divided by the global median a) Raw HindIII data. b) HindIII normalized data showing largely unbiased plot (as Iij are approximately 1). c) Raw NcoI data. d) NcoI normalized data showing largely unbiased plot (as Iij are approximately 1).

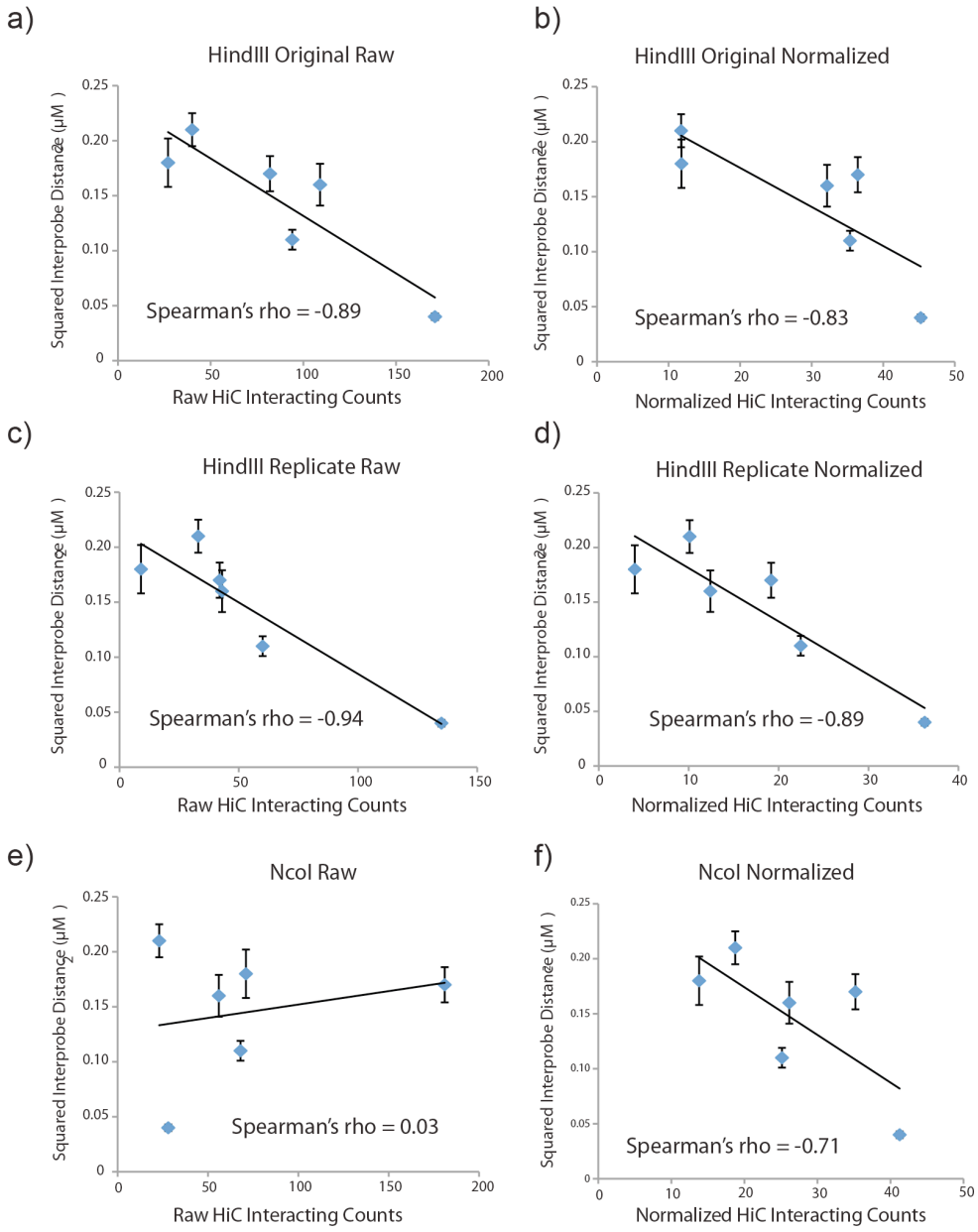

Figure 2-4: Normalized Hi-C data correlates with previously published FISH results.

a) and b) HindIII original raw and normalized data respectively. c) and d) HindIII replicate raw and normalized data respectively. e) and f) Nco1 raw and normalized data respectively. As Hi-C counts are inversly proportional to spatial distances, we expect a negative correlation among Hi-C and FISH results. While HindIII datasets show negative correlation before and after normalization, NcoI result is phenominally improved after normalization.

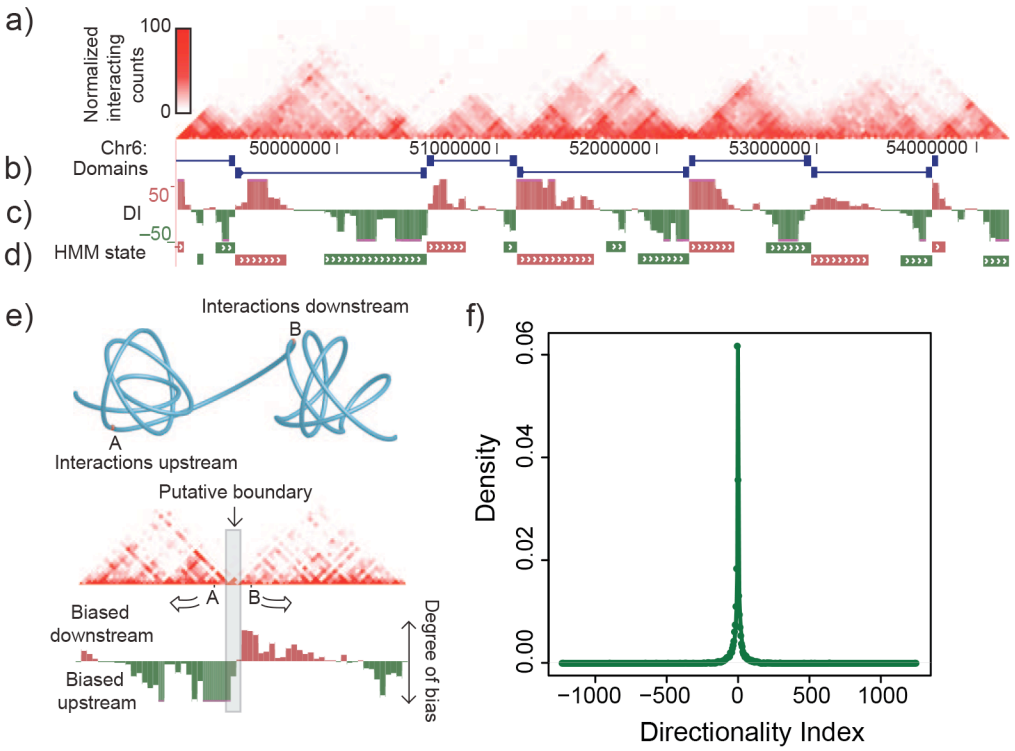

Figure 2-5: Identification of Topological domains.

a) Normalized Hi-C interaction frequencies displayed as a two-dimensional heat map, demonstrating self interacting traingles or topological domains. b) Topological domain identified from HMM state calls. c) Chi-Sequared based Directionality Index (DI) estimates used by HMM to identify topological domains. d) HMM state calls used to infer domains. For both directionality index and HMM state calls, downstream bias (red) and upstream bias (green) are indicated. e) Schematic illustrating topological domains and resulting directional bias. f) Density distribution of DIs.

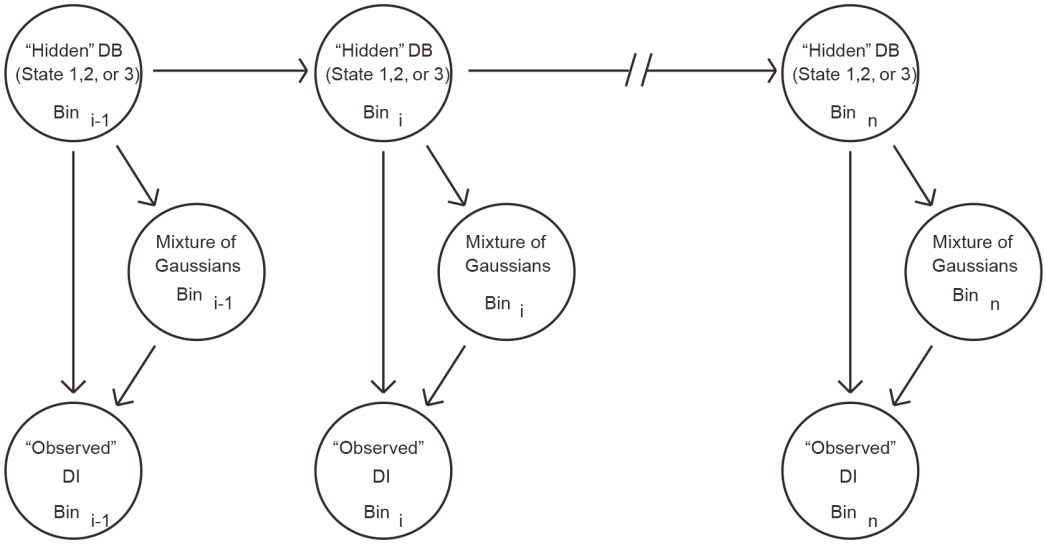

Figure 2-6: HMM with mixture of Gaussian model.

Each 40kb bin i along a chromosome having n bins has an DI value which is observed from Hi-C data. The true directionality biases are hidden and have states 1, 2, or or 3 for simplicity). Assuming that the observed DI's are a mixture of Gaussians, we determine the true directionality bias hidden state (1, 2 or 3) at bin i.

Figure 2-7: Overlap of Topological domain boundaries between Hi-C replicates.

Venn-diagrams showing high degree of overlap between boundaries called by the HMM from each pair of Hi-C replicates.

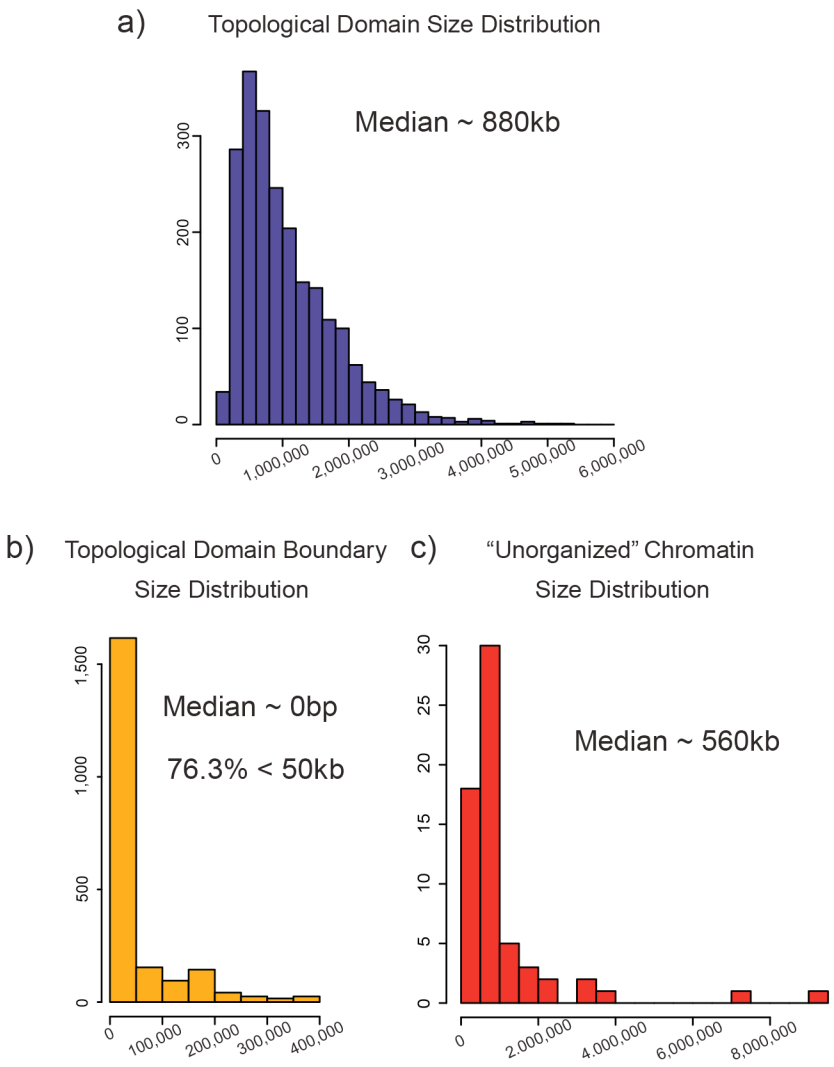

Figure 2-8: Size distribution of topological domains, boundaries, and unorganized chromatin.

a-c, Histograms of sizes of topological domains (a), topological boundaries (b), and unorganized chromatin (c). While domains are megabase long, boundary definitions are precise with 0 bp.

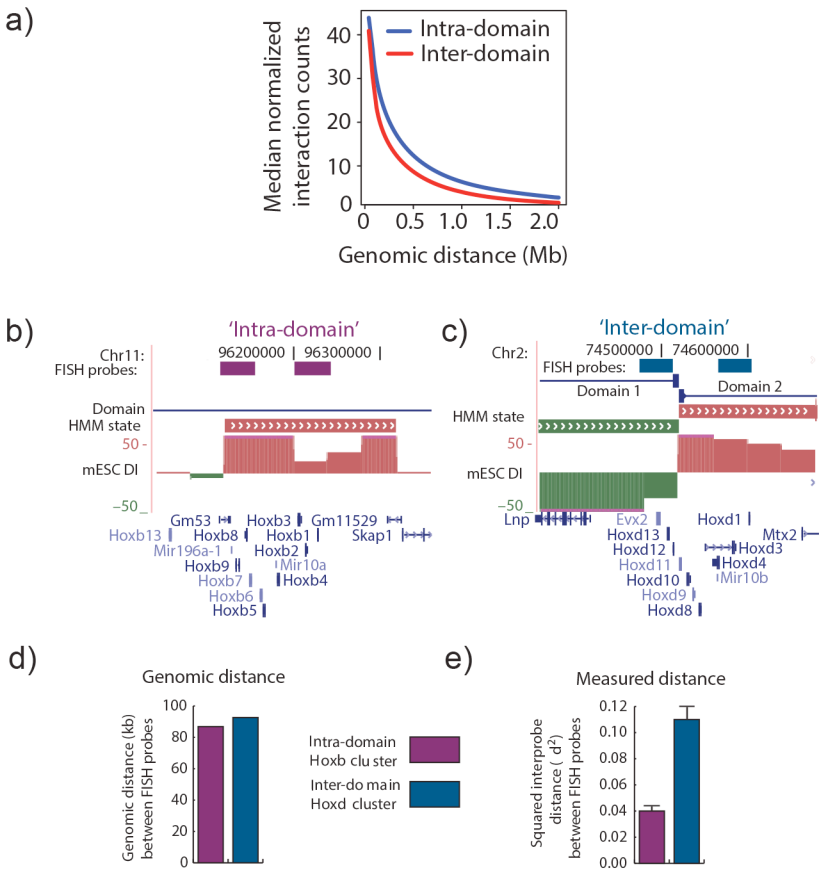

Figure 2-9: Intra-domain intercations are more frequent than inter-domain.

a) Mean interaction frequencies at all genomic distances between 40 kb to 2 Mb. Above 40 kb, the intra- versus inter- domain interaction frequencies are significantly different ($P < 0.005$, Wilcoxon test). b–e, Diagram of intra- domain (b) and inter-domain FISH probes (c) and the genomic distance between pairs (d). e, Bar chart of the squared inter-probe distance (from ref. 12) mouse ES FISH probe pairs. Error bars indicate standard error (n = 100 for each probe pair).

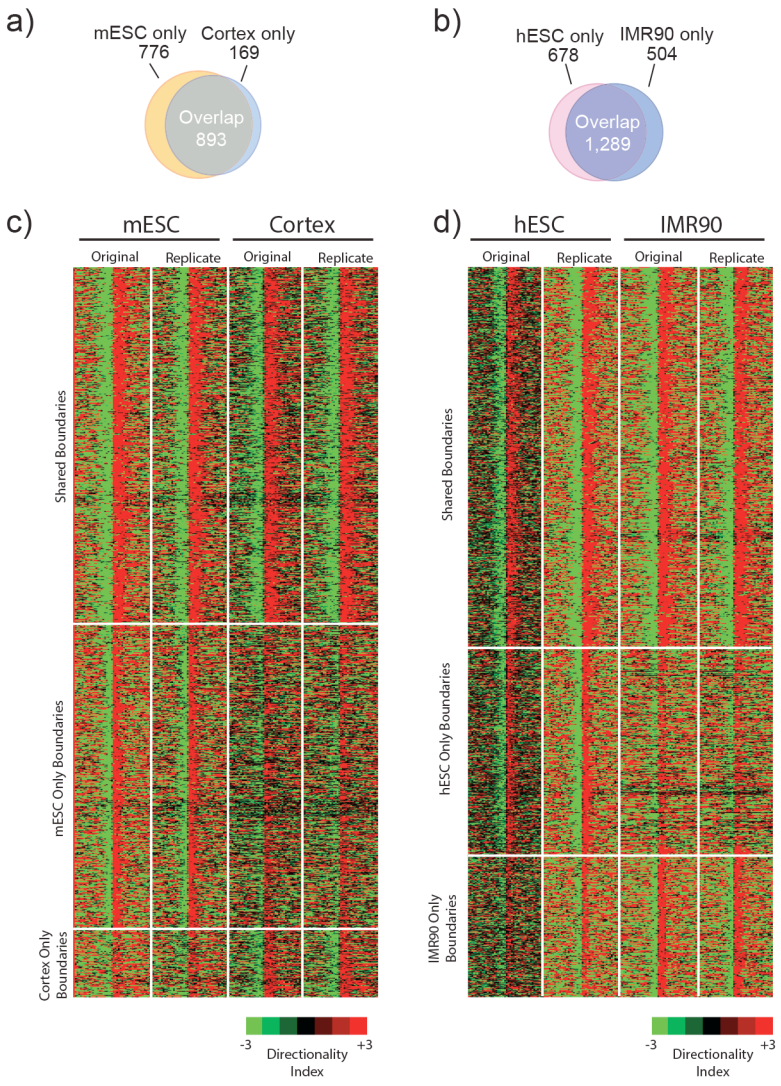

Figure 2-10: Topological domain boundaries are invariant among cell-types

a) Overlap in boundaries between mouse ES and cortex cell-types. b) Overlap in boundaries between human ES and IMR90. In both these cases, we observe a high degree of overlap among the two cell-types. c-d) Heat maps showing the directionality index surrounding the topological boundary regions. The heat maps are divided into three regions. Shared boundaries, boundaries called in cell type A and boundaries called in cell type B for mouse cells (c) and human cells (d).

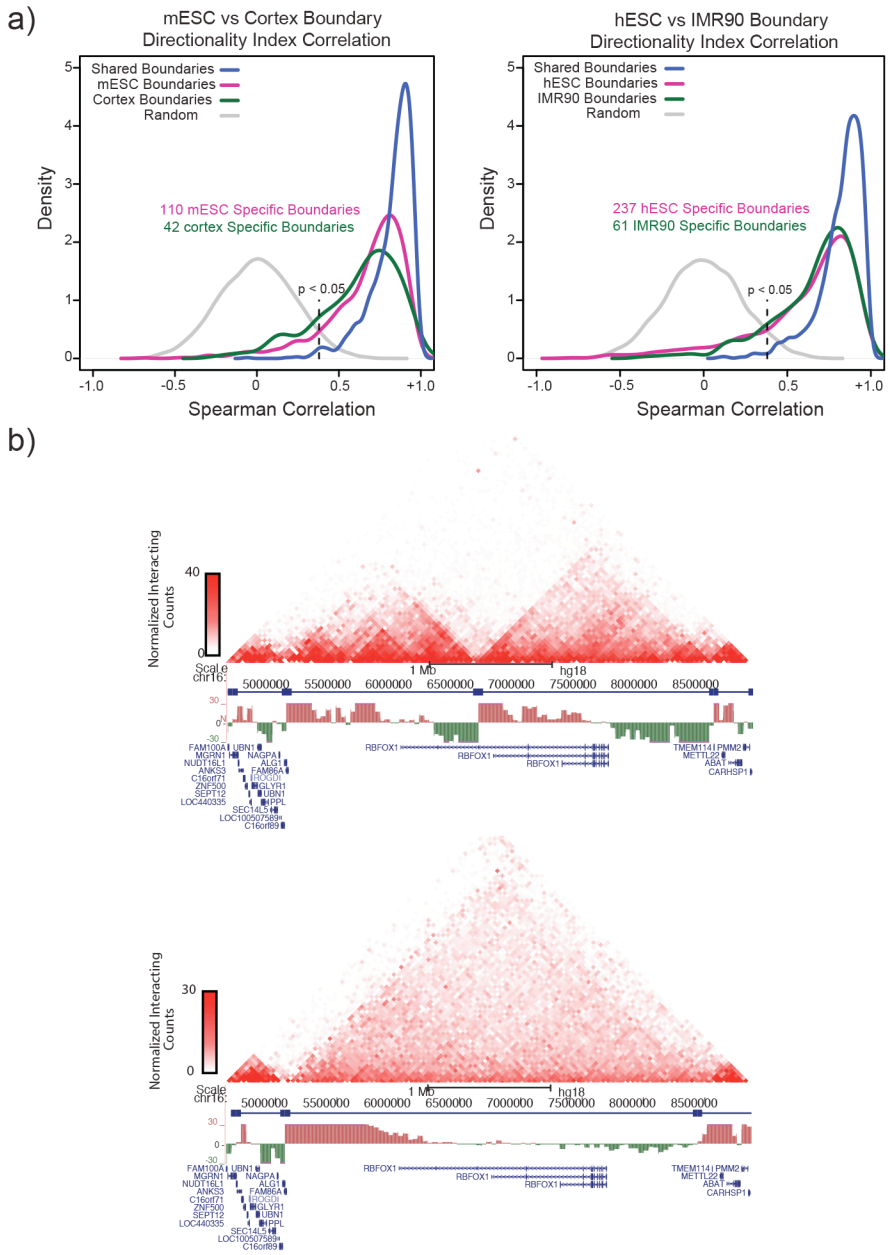

Figure 2-11: Rare cell-type specific domains.

a) Cell type specific is called if the boundary is identified by HMM in only one cell type and the spearman correlation of the directionality index is not significant when compared to a random distribution of spearman correlations. A minority of boundaries are actually called as cell types specific. b) A genome browser shot of a cell type specific domain on chromosome 16. The domain is called in hESCs and is not called in IMR90.

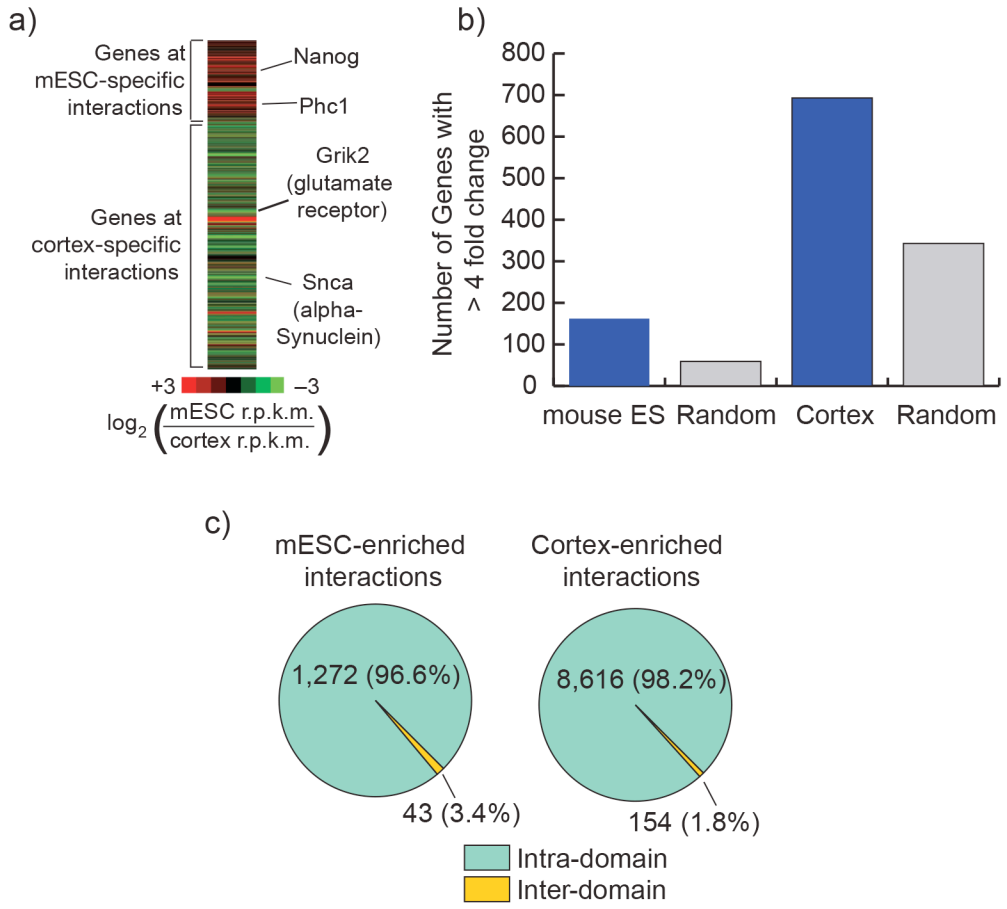

Figure 2-12: Topological domain conformation changes among cell-types leading to dynamic physical interactions and consequently differential gene regulation.

a) Heat map of the gene expression ratio between mouse ES cell and cortex of genes at dynamic interactions. b) The number of genes with > 4-fold change in gene expression that are found in a dynamic interacting region in either mouse ES cell or cortex. Shown in grey is the number of > 4-fold changed gene expected using randomly permuted dynamic interacting regions. c) Pie chart of intra-domian and inter-domain dynamic interactions.

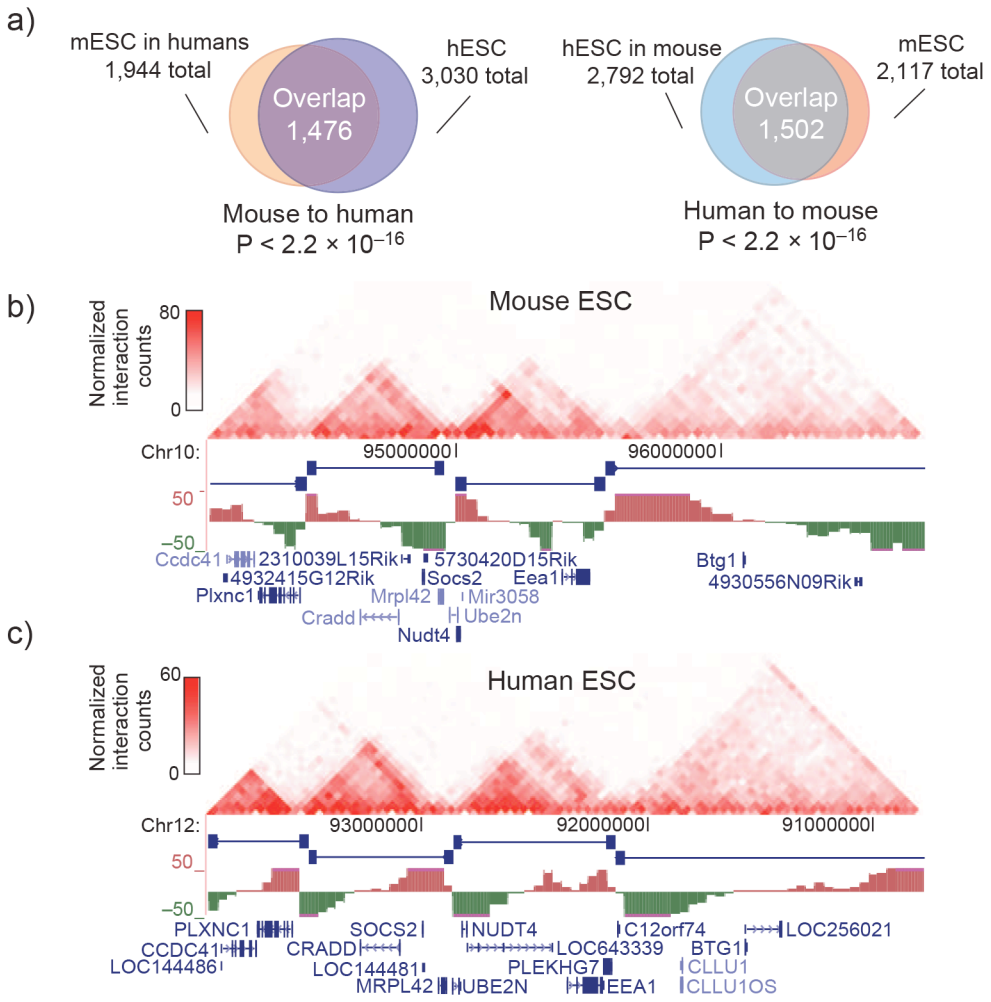

Figure 2-13: Topological domains are evolutionarily conserved accross human and mouse.

a) Overlap of boundaries between syntenic mouse and human sequences.
b-c) Genome browser shots showing domain structure over a syntenic region in the mouse (b) and human (c) ES cells.
Note: the region in humans has been inverted from its normal UCSC coordinates for proper display purposes.

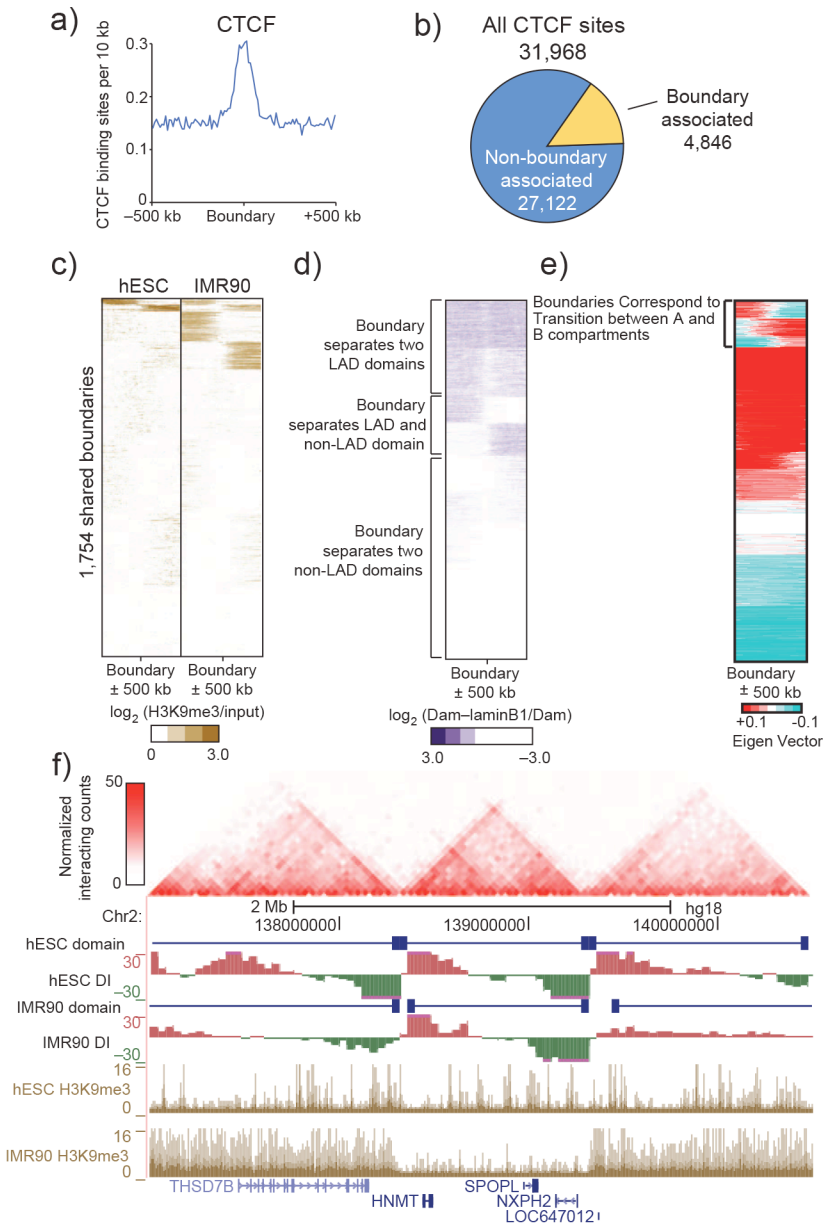

Figure 2-14: Topological Domain boundaries mark insulator/barrier elements.

a) Enrichment of CTCF at boundary regions. b) The portion of CTCF sites that are considered ‘associated’ with a boundary (within +/-20-kb). c) Heat maps of H3K9me3 at boundary sites in human ES and IMR90. d) Heat map of LADs (from ref. 27,28) surrounding the boundary regions. e) Heat map of the Eigen Vector values used to to determine the A and B compartments in mouse ES cells. f) UCSC Genome Browser shot showing heterochromatin spreading in the human ES cells (hESC) and IMR90 cells.

Acknowledgements

Chapter 2, in full, is a reprint of the material published in Nature 2012. Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, Bing Ren. “Topological domains in mammalian genomes identified by analysis of chromatin interactions”, Nature 485 (7398), 376-380, 2012. The dissertation author was a primary investigator and author of this paper. In particular, the dissertation author developed a computational method to identify topological domains and performed analyses to characterize these domains, such as their stability among cell-types and their conservation across species.

References

1. Kosak, S.T. & Groudine, M. Form follows function: The genomic organization of cellular differentiation. Genes & development 18, 1371-1384 (2004).

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

3. Cremer, T., Cremer, M., Dietzel, S., Muller, S., Solovei, I. & Fakan, S. Chromosome territories--a functional nuclear landscape. Current opinion in cell biology 18, 307-316 (2006).

4. Munkel, C., Eils, R., Dietzel, S., Zink, D., Mehring, C., Wedemann, G., Cremer, T. & Langowski, J. Compartmentalization of interphase chromosomes observed in simulation and experiment. Journal of molecular biology 285, 1053-1065 (1999).

5. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. Nature 489, 109-113 (2012).

6. Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W. & Duboule, D. The dynamic architecture of Hox gene clusters. Science 334, 222-225 (2011).

7. Jhunjhunwala, S., van Zelm, M.C., Peak, M.M., Cutchin, S., Riblet, R., van Dongen, J.J., Grosveld, F.G., Knoch, T.A. & Murre, C. The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. Cell 133, 265-279 (2008).

8. Clowney, E.J., LeGros, M.A., Mosley, C.P., Clowney, F.G., Markenskoff-Papadimitriou, E.C., Myllys, M., Barnea, G., Larabell, C.A. & Lomvardas, S. Nuclear aggregation of olfactory receptor genes governs their monogenic expression. Cell 151, 724-737 (2012).

9. Capelson, M. & Corces, V.G. Boundary elements and nuclear organization. Biology of the cell / under the auspices of the European Cell Biology Organization 96, 617-629 (2004).

10. Yokota, H., van den Engh, G., Hearst, J.E., Sachs, R.K. & Trask, B.J. Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus. The Journal of cell biology 130, 1239-1249 (1995).

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

12. Eskeland, R., Leeb, M., Grimes, G.R., Kress, C., Boyle, S., Sproul, D., Gilbert, N., Fan, Y., Skoultchi, A.I., Wutz, A. & Bickmore, W.A. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. Molecular cell 38, 452-464 (2010).

13. Bau, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J. & Marti-Renom, M.A. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. Nature structural & molecular biology 18, 107-114 (2011).

14. Chambeyron, S. & Bickmore, W.A. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. Genes & development 18, 1119-1130 (2004).

15. Solovei, I., Cavallo, A., Schermelleh, L., Jaunin, F., Scasselati, C., Cmarko, D., Cremer, C., Fakan, S. & Cremer, T. Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). Experimental cell research 276, 10-23 (2002).

16. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. Science 295, 1306-1311 (2002).

17. Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V. & Ren, B. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116-120 (2012).

18. Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., Wysocka, J., Lei, M., Dekker, J., Helms, J.A. & Chang, H.Y. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472, 120-124 (2011).

19. Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., Taatjes, D.J., Dekker, J. & Young, R.A. Mediator and cohesin connect gene expression and chromatin architecture. Nature 467, 430-435 (2010).

20. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature genetics 43, 1059-1065 (2011).

21. Scott, K.C., Taubman, A.D. & Geyer, P.K. Enhancer blocking by the Drosophila gypsy insulator depends upon insulator anatomy and enhancer strength. Genetics 153, 787-798 (1999).

22. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. & Haussler, D. The human genome browser at UCSC. Genome research 12, 996-1006 (2002).

23. Phillips, J.E. & Corces, V.G. CTCF: master weaver of the genome. Cell 137, 1194-1211 (2009).

24. Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F., Wong, E., Sheng, J., Zhang, Y., Poh, T., Chan, C.S., Kunarso, G., Shahab, A., Bourque, G., Cacheux-Rataboul, V., Sung, W.K., Ruan, Y. & Wei, C.L. CTCF-mediated functional chromatin interactome in pluripotent cells. Nature genetics 43, 630-638 (2011).

25. Xie, W., Barr, C.L., Kim, A., Yue, F., Lee, A.Y., Eubanks, J., Dempster, E.L. & Ren, B. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell 148, 816-831 (2012).

26. Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S., Antosiewicz-Bourget, J., Ye, Z., Espinoza, C., Agarwahl, S., Shen, L., Ruotti, V., Wang, W., Stewart, R., Thomson, J.A.,

Ecker, J.R. & Ren, B. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell stem cell 6, 479-491 (2010).

27. Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W., Solovei, I., Brugman, W., Graf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., Reinders, M., Wessels, L. & van Steensel, B. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Molecular cell 38, 603-613 (2010).

28. Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W. & van Steensel, B. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453, 948-951 (2008).

29. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. & Cavalli, G. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 148, 458-472 (2012).

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

31. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Bluthgen, N., Dekker, J. & Heard, E. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381-385 (2012).

32. Jinek, M., East, A., Cheng, A., Lin, S., Ma, E. & Doudna, J. RNA-programmed genome editing in human cells. eLife 2, e00471 (2013).

33. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. & Church, G.M. RNA-guided human genome engineering via Cas9. Science 339, 823-826 (2013).

34. Cade, L., Reyon, D., Hwang, W.Y., Tsai, S.Q., Patel, S., Khayter, C., Joung, J.K., Sander, J.D., Peterson, R.T. & Yeh, J.R. Highly efficient generation of heritable zebrafish gene mutations using homo- and heterodimeric TALENs. Nucleic acids research 40, 8001-8010 (2012).

35. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. & Charpentier, E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337, 816-821 (2012).

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

37. Fullwood, M.J. & Ruan, Y. ChIP-based methods for the identification of long-range chromatin interactions. Journal of cellular biochemistry 107, 30-39 (2009).

38. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. & Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106, 9362-9367 (2009).

39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760 (2009).

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

41. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. & Liu, X.S. Model-based analysis of ChIP-Seq (MACS). Genome biology 9, R137 (2008).

Chapter 3: Repurposing Hi-C towards generating haplotypes

Abstract

Rapid advances in high-throughput sequencing facilitate variant discovery and genotyping, but linking variants into a single haplotype remains a challenge. Here I demonstrate HaploSeq, a novel approach for assembling chromosome-scale haplotypes that exploits the existence of ‘chromosome territories’. Our lab performed Hi-C and I show that alleles on homologous chromosomes occupy distinct territories, and therefore this experimental protocol preferentially recovers physically linked DNA variants on a homolog. Computational analysis of such data sets allows for accurate (~99.5%) reconstruction of chromosome-spanning haplotypes for ~95% of alleles in hybrid mouse cells with 30× sequencing coverage. To resolve haplotypes for a human genome, which has a low density of variants, I coupled HaploSeq with local conditional phasing to obtain haplotypes for ~81% of alleles with ~98% accuracy from just 17× sequencing. Whereas Hi-C was originally designed to investigate spatial organization of the genome, I have repurposed it as a general tool for haplotyping.

Introduction

Rapid progress in DNA shotgun sequencing technologies has enabled systematic identification of the genetic variants of an individual[1–4]. However, as the human genome consists of two homologous sets of chromosomes, understanding the true genetic makeup of an individual requires delineation of the maternal and paternal copies or haplotypes of the genetic material. Obtaining

a haplotype in an individual is useful in several ways. First, haplotypes are useful clinically in predicting outcomes for donor-host matching in organ transplantation[5,6] and are increasingly used as a means to detect disease associations[7–9]. Second, in genes that show compound heterozygosity, haplotypes provide information as to whether two deleterious variants are located on the same allele, greatly affecting the prediction of whether inheritance of these variants is harmful[10–12]. Third, haplotypes from groups of individuals have provided information on population structure[13–15] and the evolutionary history of the human race[16]. Lastly, recently described widespread allelic imbalances in gene expression suggest that genetic or epigenetic differences between alleles may contribute to quantitative differences in expression[17–20]. An understanding of haplotype structure will therefore be critical for delineating the mechanisms of variants that contribute to allelic imbalances. Taken together, knowledge of complete haplotype structure in individuals is essential for advancing personalized medicine.

Recognizing the importance of haplotypes, several groups have sought to expand our understanding of haplotype structures at the level of both populations and individuals. Initiatives such as the International Hapmap Project[13] and the 1000 Genomes Project[14,15] have attempted to systematically reconstruct haplotypes through linkage disequilibrium measures based on populations of unrelated individuals. However, the average length of accurately phased haplotypes generated using this approach is limited to ~300 kb[21,22]. Alternatively, genotyping parent-child trios can determine whole-genome haplotypes in the

child, but such methods are constrained by their higher cost and the sample availability of the two biological parents.

Numerous experimental methods have also been developed to facilitate direct haplotype phasing of an individual, including long-fragment-read sequencing[23], mate-pair sequencing[24], fosmid sequencing[4,25–27] and dilution-based sequencing[28]. At best, these methods can reconstruct haplotypes ranging from several kilobases to about a megabase, but none can achieve chromosome-spanning haplotypes. Whole-chromosome haplotype phasing has been achieved by sequencing based on fluorescence-activated cell sorting[29], chromosome-segregation followed by sequencing[21] and chromosome microdissection–based sequencing[30]. However, these methods only phase a fraction of the heterozygous variants in an individual, and more importantly, they are technically challenging to perform or require specialized instruments. Recently, whole-genome haplotyping has been performed using genotyping from sperm cells[31]. However, this approach is not applicable to the general population and requires the deconvolution of complex meiotic recombination patterns.

Computational analysis has shown that an important factor in haplotype reconstruction from DNA shotgun sequencing methods is the length of the sequenced genomic fragment[32]. For example, longer haplotypes can be obtained using mate-pair sequencing (fragment or insert size, ~5 kb) compared with conventional genome sequencing (fragment or insert size ~500 bp) (Fig. 3-1a). However, it is technically difficult to isolate and sequence DNA fragments that are longer than what is already obtained using fosmid clones. Hence, using existing

shotgun sequencing approaches, it is difficult to generate haplotype blocks longer than 1 million bases, even at ultra-deep sequencing coverage (Fig. 3-1b).

Here I describe an approach, termed HaploSeq, for haplotyping by combining Hi-C[33–35] with a probabilistic algorithm for haplotype assembly[24]. We have experimentally validated HaploSeq in a hybrid mouse embryonic stem cell line and a human lymphoblastoid cell line in which the complete haplotypes were known a priori. With HaploSeq, chromosome-spanning haplotype reconstruction can be achieved with >95% of alleles linked at an accuracy of ~99.5% in mouse. In the human cell line, I coupled HaploSeq with local conditional phasing to obtain chromosome-spanning haplotypes at ~81% resolution with an accuracy of ~98% using just 17× coverage of genome sequencing. These results establish the utility of Hi-C for haplotyping in human populations.

Results

Experimental strategy of HaploSeq

In HaploSeq, we first perform the Hi-C protocol[34]. As this method captures DNA fragments from two distant genomic loci that looped together in three-dimensional space in vivo[33–35], sequencing of the resulting DNA library generates reads having ‘insert sizes’ ranging from several hundred base pairs to tens of millions of base pairs (Fig. 3-2a). Thus, although the short DNA fragments generated in a Hi-C experiment can yield small haplotype blocks, long fragments ultimately can link these small blocks together (Fig. 3-2b). With enough

sequencing coverage, such an approach has the potential to link variants in discontinuous blocks and assemble every such block into a single haplotype.

One complicating factor is that Hi-C can capture interactions both in cis within an individual allele and in trans between homologous and non-homologous chromosomes. Although non-homologous trans interactions between different chromosomes do not affect phasing, interactions in trans between homologous chromosomes (referred to as h-trans hereafter) might complicate haplotype reconstruction if h-trans interactions were as frequent as cis interactions. Therefore, I set out to determine the relative frequency of h-trans versus cis interactions in Hi-C sequencing data. To accomplish this, our lab performed Hi-C with 30× sequencing coverage in a hybrid mouse embryonic stem (ES) cell line derived from a cross between two inbred homozygous strains (Mus musculus castaneous (CAST) and 129S4/SvJae (J129)), which were previously sequenced (Methods). Owing to its homozygous nature, the maternal and paternal haplotypes are known a priori, and the frequency of interactions between alleles can then be explicitly tested.

To determine the extent of intrahaplotype (cis) versus interhaplotype (h-trans) interactions, we used the prior haplotype information to distinguish reads from CAST and J129 alleles. We first visually checked the pattern of interactions between every allele, finding that the CAST and J129 alleles for each chromosome were largely self-interacting and distinct (Fig. 3-3). Such a pattern has been previously observed in Hi-C studies and is analogous to the long-established concept that chromosomes occupy distinct, self-associated

territories, known as “chromosome territories,” within the interphase nucleus[34,35]. However, previous Hi-C studies did not distinguish whether the two alleles for a given chromosome also occupy distinct, individual, chromosome territories[34,35]. Overall, we observed 2% h-trans interactions among the total reads originating and ending on the two homologous chromosomes (Fig. 3-4a). In addition, the probability of a DNA read being in h-trans versus in cis appears to increase as a function of the insert size between the read pairs (Fig. 3-4b). Because of this trend, I capped the maximum insert size of Hi-C reads at 30 Mb to reduce the overall number of h-trans interactions to ~0.6% (Fig. 3-4c). Currently, we cannot determine if these rare h-trans interactions are due to noise in the data or to biological phenomena, such as homologous pairing of chromosomes[36]. Regardless, these observations indicate that h-trans interactions are rare, a prerequisite for HaploSeq analysis to succeed.

Predicting accurate chromosome-span haplotypes in mouse

Rare h-trans interacting reads and phenomena such as sequencing errors at the variant locations can cause erroneous connections between homologous chromosomes and complicate the reconstruction of haplotypes. To overcome these problems, I incorporated HapCUT[24] software into HaploSeq analysis to probabilistically predict haplotypes. Because Hi-C generates larger graphs than conventional genome sequencing or mate-pair sequencing, we modified HapCUT to balance computing time and number of iterations, so that the

haplotypes can be predicted with reasonable speed and high accuracy (Methods).

To test the ability of HapCUT to generate haplotype blocks, I used the CAST×J129 mouse Hi-C data. I allowed HapCUT to reconstruct de novo haplotype blocks of the heterozygous variants and used the metrics of completeness, resolution and accuracy to assess the performance of HaploSeq (Fig. 3-5). To assess completeness, I analyzed the span of the haplotype blocks generated for each chromosome. I observed that each chromosome contains one block with the most heterozygous variants phased (MVP) and many other small blocks. However, the MVP block is the most useful as it phases a large fraction of variants. The MVP block spanned >99.9% of the phasable base-pairs for each chromosome (Table 3-1), demonstrating that HaploSeq analysis using Hi-C data can generate complete, chromosome-spanning haplotypes.

Although completeness is defined as the base-pair span of the MVP block, resolution is defined as the fraction of phased heterozygous variants relative to the total variants spanned in the MVP block (Fig. 3-5). These MVP blocks generated for each chromosome are of high resolution, as we could phase about 95% of the heterozygous variants on any given chromosome (Table 3-1). As 99.6% of variants are covered by at least one read, the inability to link the 5% of heterozygous variants is primarily due to the inability to link heterozygous variants to the MVP haplotype block. Consequently, although the MVP block spans the majority of the chromosome, it has gaps that in total contain ~5% of the heterozygous variants.

To assess the accuracy of the heterozygous variants within the MVP block, I compared the predicted haplotypes generated de novo by HaploSeq analysis with the known haplotypes of the CAST and J129 alleles. I defined accuracy as the fraction of phased heterozygous variants that were correctly phased in the MVP block (Fig. 3-5). Of the variants that were assigned to the MVP haplotype block, I observed >99.5% accuracy in distinguishing between the two known haplotypes (Table 3-1). Lastly, as I had previously demonstrated that the h-trans interaction probability increases with the genomic distance separating two sequencing reads (Fig. 3-4b), I incorporated the h-trans interaction probabilities into the HapCUT algorithm (Methods) and constrained the maximum insert size to be 30 megabase. These conditions did not sacrifice the completeness of the haplotypes we generated. Instead, I observed a further improvement in the accuracy of the variants in the MVP block with a modest reduction of the resolution of the variants phased (Fig. 3-6a,b). In summary, these results demonstrate that HaploSeq analysis yields complete, high-resolution and accurate haplotypes for all autosomes.

Previous haplotyping efforts have often combined different shotgun sequencing methods to improve phasing. For instance, whole-genome sequencing has been combined with mate-pair sequencing[24]. To see if this approach would also improve haplotyping with proximity-ligation data, I simulated 20× coverage DNA sequencing data for conventional paired-end shotgun DNA sequencing (i.e., WGS), mate-pair sequencing, fosmids and proximity ligation. As expected, combining WGS with mate pair or fosmid data resulted in fragmented,

incomplete haplotype blocks (Fig. 3-7a,b). In contrast, performing HaploSeq analysis using Hi-C in combination with WGS data did not increase the completeness of the haplotypes generated (Fig. 3-7a) but did improve their resolution (Fig. 3-7b), suggesting that adding WGS to HaploSeq analysis may be a viable strategy in cases where the resolution of haplotypes must be maximized.

Performance of HaploSeq depends on variant density

A distinct feature of the CAST×J129 ES cell line is the high density of heterozygous variants present throughout the genome. On average, there is a heterozygous variant every 150 bases, which is 7–10 times more frequent than in humans[1,2]. As a first test of the feasibility of using HaploSeq to generate haplotypes in human cells, I sub-sampled heterozygous variants in the CAST×J129 data so that the variant density mimics that in human populations. I then tested how lower variant density affects the ability of HaploSeq to reconstruct haplotypes. Although lower variant density did result in fewer usable reads (Fig. 3-8a,b), I still observed complete haplotypes over each chromosome with only a marginal decrease in accuracy (from ~99.6% to ~99.2%, Table 3-2). However, the MVP block generated using a variant density similar to that observed in the human genome had a lower resolution. Approximately 32% of heterozygous variants were phased in the MVP block (Table 3-2), instead of 95% in the high-density case (Table 3-1). In summary, a low density of variants does not affect completeness or accuracy, but does substantially affect the resolution of chromosome-spanning haplotypes by HaploSeq analysis.

HaploSeq analysis of a human individual

To realistically assess the ability of HaploSeq to phase haplotypes in humans, our lab performed Hi-C at ~17× coverage on the GM12878 lymphoblastoid cell line. The 1000 Genomes Project has previously inferred the complete haplotype of this cell line from whole-genome sequencing of parent-child trio[14]. HaploSeq generated chromosome-spanning haplotypes in all chromosomes of the GM12878 cells (Table 3-3). Of note, previous methods attempting haplotype reconstruction in humans have been unable to reconstruct haplotypes spanning across the highly repetitive centromeric regions of metacentric chromosomes[4,23,25–28]. Using HaploSeq, I generated haplotypes that accurately spanned the centromere in all metacentric chromosomes with the exception of chromosome 9, where an erroneous linkage caused switching of haplotype calls at the centromere (Fig. 3-9). Chromosome 9 has both a large 15-Mbp, poorly mapped centromere region and relatively lowers usable coverage (13.7×). I hypothesized that additional coverage might offer us a better chance in accurately spanning the centromere. Therefore, I combined our Hi-C data with previously generated Hi-C and tethered chromosome confirmation capture (TCC) data. TCC is a Hi-C variant using solid support ligation[35] that generates similar data as a Hi-C experiment with slightly better ability to capture long-range chromatin interactions (Fig. 3-10). Using this combined data set, I increased the coverage of chromosome 9 to ~15×, which allowed accurate phasing of the entire chromosome. In summary, I generated complete chromosome-spanning

haplotypes for all human chromosomes including chromosome X, albeit at reduced resolution of ~22% (Table 3-3).

Combining HaploSeq and local conditional phasing

Although I generated chromosome-spanning haplotypes using HaploSeq, I was unable to achieve a high resolution of variants phased owing to the low variant density in the human population. I reasoned that the gaps in the MVP block containing unphased variants could be probabilistically linked to the MVP block using linkage disequilibrium patterns derived from population-scale sequencing data. For this purpose, I used the HaploSeq-generated, chromosome-spanning haplotype as a 'seed haplotype' to guide the local phasing using the Beagle (v4.0)[37] software and sequencing data from the 1000 Genomes Project[15].

To initially assess the effectiveness of this approach, I simulated chromosome-spanning seed haplotypes in the GM12878 genome with different percentages of variants phased in the MVP block. My simulation results suggest that I can accurately infer local phasing even at low-resolution seed haplotype inputs (3% error at 10% seed haplotype resolution; Fig. 3-11a). Owing to complex population structures, occasional mismatches occurred between phase predictions from local haplotypes predicted by Beagle and the HaploSeq-generated seed haplotype. To correct these mismatches, I filtered heterozygous variants with <100% agreement with the seed haplotype in a local neighborhood window surrounding the heterozygous variant. This filtering reduced the error

rate to ~0.7% regardless of seed haplotype resolution (Fig. 3-11a). Consequently, the fraction of heterozygous variants for which I can infer local phasing increased with greater seed haplotype resolution (Fig. 3-11a). By contrast, altering the neighborhood window size did not substantially increase accuracy (Fig. 3-11b).

Encouraged by these results, I used the MVP chromosome-spanning haplotypes generated from HaploSeq analysis as seed haplotypes and performed local conditional phasing. Overall, I generated chromosome-spanning haplotypes with ~81% resolution at an average accuracy of ~98% (Table 3-4). Therefore, by coupling HaploSeq analysis and local conditional phasing, I achieved high-resolution and accurate chromosome-spanning haplotypes in humans.

Sequencing requirements for obtaining haplotypes

From my local conditional phasing analysis, it seems that a seed haplotype with ~20–30% resolution is sufficient to obtain accurate and high-resolution, chromosome-spanning haplotypes. A subsequent question therefore is, what are the minimal experimental requirements to achieve chromosome-spanning seed haplotypes with ~20–30% resolution? To investigate this, I simulated Hi-C data with varying read lengths and sequencing coverage. Based on the simulation, achieving chromosome-spanning haplotypes depends on obtaining a usable sequencing coverage of ~15× for most of the read lengths tested (Fig. 3-12a). However, chromosome-spanning seed haplotypes alone are

not enough for achieving high-resolution haplotypes through local conditional phasing. In particular, the resulting sparse seed haplotype graph may limit the ability to generate final high-resolution haplotypes. To increase the resolution of the seed haplotype once complete seed haplotypes are obtained, one must increase coverage, either through higher sequencing depth or longer read lengths (Fig. 3-12b). I observed that 50- to 100-bp paired-end reads balanced completeness and resolution, and achieved the desired fraction of ~20–30% resolution at ~25–30× usable coverage.

Discussion

I describe a strategy to reconstruct chromosome-spanning haplotypes for an individual. Although the density of heterozygous variants contributes strongly to the resolution of the generated haplotypes, I showed that this complication could be resolved by using local conditional phasing from population data15 (Fig. 3-13). Compared with other haplotyping approaches that can reconstruct complete haplotypes[21,29,30], HaploSeq is the most suitable for a clinical and laboratory setting, where reagents and equipment required are readily available. Furthermore, HaploSeq is more widely applicable than approaches based on sperm cell genotyping[31], as it can generate whole-genome haplotypes from intact cells of any individual or cell line.

We anticipate that HaploSeq will be useful for personalized medicine. Determination of haplotypes in individuals has the potential to reveal novel

haplotype-disease associations, some of which have already been identified on smaller scales[38–40]. In addition, complete haplotypes will be essential for understanding allelic biases in gene expression, which will contribute to knowledge of genetic and epigenetic polymorphisms in the population and their phenotypic consequences at a molecular level[17–20]. As a result, whole-genome haplotyping has applications across several fields, such as pharmacogenomics, genetic diagnostics, agricultural crop breeding and genetic engineering of animals.

Hi-C was originally invented to study the spatial organization of chromosomes[34]. Here we show that it is also valuable for studying the genetic makeup of an individual. In principle, Hi-C data can also be used for genotyping, along the same lines as WGS. Although variants far from restriction enzyme cut sites are less likely to be genotyped owing to biases from Hi-C approach, population-based imputation[22] of variants not yet genotyped can improve the performance of genotype calling. Because all this can be done using a single experiment, HaploSeq has the potential to become a general tool for whole-genome analysis in the future.

Methods

Genotyping

Variant calls and genotypes for GM12878 were downloaded[43] and these were used for haplotype reconstruction by Hi-C. Phasing Information for GM12878 was downloaded from 1000 Genomes Project[14].

For generating genotype calls for the hybrid CAST×J129 cells, we downloaded parental genome sequencing data from publicly available databases. For CAST, we downloaded the genome sequence from the European Nucleotide Archive (accession number ERP000042). S129/SvJae genome sequencing data was downloaded from the Sequence Read Archive (accession number SRX037820). Reads were aligned to the mm9 genome using Novoalign (www.novocraft.com) and using samtools[44], and we filtered out unmapped reads and PCR duplicates. The final aligned data sets were processed using the Genome Analysis Toolkit (GATK)[45]. Specifically, we performed indel realignment and variant recalibration. The GATK Unified Genotyper was used to make single-nucleotide polymorphism (SNP) and indel calls. We filtered out variants that did not meet the GATK quality filters or that were called as heterozygous variants, as the genome sequencing was performed in homozygous parental inbred mice. The genotype calls in the parents were used both to determine the extent of interactions in cis versus h-trans to learn the phasing of hybrid CAST×J129 cells a priori to haplotype reconstruction.

Hi-C read alignment

For Hi-C read alignment, we aligned Hi-C reads to the mm9 (mouse) or the hg18 (human) genome. In each case, we masked any bases in the genome

that were genotyped as SNPs in either Mus musculus castaneus or S129/SvJae (for mouse) or GM12878 (for humans). These bases were masked to “N” in order to reduce reference bias mapping artifacts. Hi-C reads were aligned iteratively as single-end reads using Novoalign and samtools[44]. Specifically, for iterative alignment, we first aligned the entire sequencing read to either the mouse or human genome. Unmapped reads were then trimmed by 5 bp and realigned. This process was repeated until the read successfully aligned to the genome or until the trimmed read was less than 25 bp long. Iterative alignment is useful for Hi-C data because certain reads will span a proximity-ligation junction and fail to successfully align to the genome due to gaps and mismatches. Iteratively trimming unmapped reads has the potential to allow these reads to align successfully to the genome when the trimming removes the part of the read that spans the ligation junction. After iterative alignment of reads as single ends is complete, the reads are manually paired using in-house scripts. Unmapped and PCR duplicate reads are removed. The aligned data sets are then finally subjected to GATK[45] indel realignment and variant recalibration.

Usable coverage

For phasing using HapCUT, we utilize both intra-chromosomal and inter-chromosomal reads. For inter-chromosomal reads, I consider each inter-chromosomal read pair as two single-end reads, as the paired information for such reads is not useful for phasing. In contrast, all intra-chromosomal reads are considered for phasing. The probability of a single read to harbor more than one

variant is small, especially in humans where the variant density is relatively low. This, in combination with the fact that only the paired intra-chromosomal reads will have large insert sizes, means that the vast majority of reads that contribute to the success of haplotype phasing are the intra-chromosomal reads. Therefore, I define the “usable coverage” as the genomic coverage derived from intra-chromosomal reads only.

Our Hi-C experiment generated ~22% inter-chromosomal reads in CAST×J129, whereas ~55% of the reads in GM12878 were inter-chromosomal. In other words, 620 M paired-end reads out of 795 M were useful in CAST×J129, with a usable coverage of 30×. In humans, only 262 M paired-end reads out of 577 M were useful, resulting in a usable coverage of 17×. In our experience, the fraction of all reads that are intra-chromosomal versus inter-chromosomal in a Hi-C experiment may vary between experiments and across cell types.

Analysis of HaploSeq data using HapCUT

I used the HapCUT[24] algorithm to perform the computational aspects of HaploSeq, This method was originally designed to work on conventional genome sequencing (WGS) or mate-pair sequencing data. HapCUT constructs a graph with the heterozygous variants as nodes and DNA fragment(s) connecting two nodes as edges. Therefore, only fragments with at least two heterozygous variants are useful for haplotype phasing. HapCUT extracts such ‘haplotype-informative’ fragments from a coordinate-sorted BAM file using a sorting method that stores each potential haplotype-informative read in a buffer until its mate is

seen. We customized the buffer size to allow HapCUT handle large insert–sized Hi-C reads.

HapCUT uses a greedy max-cut heuristic to identify the haplotype solution for each connected component in the graph with the lowest score under the MEC scoring function. In particular, the original HapCUT algorithm used O(n) iterations to find the best cut. Because Hi-C data resulted in chromosomal spanning haplotypes with a single large connected component, the default method took several days of computing time to phase the CAST×J129 genome. To reduce the computation time, I assessed the impact of reducing the number of max-cut iterations on the accuracy of phasing. For CAST×J129 system, increasing the number of max-cut iterations beyond 1,000 did not significantly improve the accuracy. For GM12878, I allowed up to 100,000 iterations.

Once a best-cut solution is achieved, that solution is iterated multiple times to improve upon the current best-cut solution among other possible best cuts in the solution space. I used a maximum of 21 such iterations in CAST×J129 and 101 in GM12878 cells. My parameters in GM12878 cells allowed HapCUT to obtain higher accuracy given the lower variant density and reduced sequence coverage compared to the mouse data. The modified version of HapCUT can be downloaded from https://sites.google.com/site/vibansal/software/hapcut.

Maximum insert size analysis

As previously mentioned the probability of a Hi-C read being in cis versus h-trans varies as a function of the distance between the two read pairs (Fig. 2c).

At shorter genomic distances, the probability that an intrachromosomal read is in h-trans is very low. At large distances (>30 Mbp), this probability rises substantially and is in theory more likely to introduce erroneous connections for HapCUT to phase. To account for this, I used the Hi-C data for chromosomes 1, 5, 10, 15 and 19 in the CAST×J129 data and repeated haplotype reconstruction allowing variable maximum insert size values. I excluded any reads where the insert size between reads was greater than the allowable maximum insert size. I performed this analysis using the low variant density case as lower density was most amenable for applications in humans. This step resulted in increase in accuracy of HaploSeq analysis with moderate reduction in resolution.

Insert size–dependent probability correction

A useful feature of the HapCUT algorithm is that it accounts for the base quality score at a variant location to calculate the score of a potential haplotype. In other words, if a sequencing read that links two variants and the base quality at one variant location is low, this read is given relatively lower weight by HapCUT in generating its final haplotype calls. Therefore, HapCUT can use this information to try to disregard potential sequencing errors from making erroneous haplotype connections. As we previously mentioned, in Hi-C data errors may also arise due to h-trans interactions, which are much more frequent than sequencing errors and show a distance-dependent behavior. Therefore, I attempted to account for the likelihood of an interaction being in cis versus h-trans based on the distance between the two reads. I used the CAST×J129 Hi-C data to identify

reads that are in cis or h-trans. I binned the insert sizes into 50-kb bins and estimated the probability of a read being h-trans (#h-trans/(#cis+#h-trans). I then used local regression (LOWESS) at 2% smoothing to predict h-trans probabilities at any given insert size. For every intrachromosomal read, I multiplied the cis probabilities (1 – h-trans) with the base qualities to account for the odds of this intrachromosomal read being a h-trans interaction. As a result, reads that are more likely to be h-trans are given lower weight by HapCUT in identifying the haplotype solution.

Adding h-trans interaction probabilities increases HaploSeq accuracy moderately, without having any affect on resolution. As a comparison, maximum insert size of 30 Mb had an error rate of 1.1% in chromosome 19. After adding h-trans probabilities, the error rate is 0.9%, where error rate is defined as 1 – accuracy.

Local conditional phasing simulation

In order to study our ability to perform local phasing at different percentages of resolution, I performed a stepwise analysis. First, I generated seed haplotypes at different resolutions. Then, I used Beagle (v4.0)[37] to perform local phasing under the guidance of the seed haplotype. Finally, I checked accuracy of local phasing by comparing it to phasing information known a priori from 1000 Genomes Project.

To simulate seed haplotypes at different resolutions, I first simulated seed genotypes. I used different combinations of read length and coverage to obtain

seed genotypes of various resolutions. In particular, I used Hi-C intra-chromosomal read starting positions from H1 and H1-derived cells (unpublished data) to generate pairs of reads of a given read length and coverage. This allowed us to maintain the Hi-C data structure and the observed distribution of insert sizes in the simulated data. To generate the seed genotype, I constructed a graph with nodes representing heterozygous variants in GM12878 (chromosome 1) and edges corresponding to reads that cover multiple variants. This graph is essentially a genotype graph because we don't know the phasing yet. Hence, the whole point of this graph is to provide a two subset of variants: one that is a part of the seed genotype and other that is not (which are the gaps to be inferred by local phasing), based on the resolution and Hi-C data structure. I generated seed genotypes at required parameters of read length and coverage to attain a specific resolution. I used these seed genotypes for both local phasing and to study the minimal requirements for generating seed haplotypes of enough resolution. These two analyses were done independently and in both cases, I repeated generating seed genotypes and downstream analysis ten times to note the average results.

To perform local conditional phasing, I need an a priori haplotype system to check accuracy of our local conditional phasing. Because a priori haplotype information from the trio covers only a fraction of heterozygous variants, I decided to perform local phasing simulation only on the trio subset. Specifically, I required every variant that was part of either seed genotype or "gaps" to be part of the 1000 Genomes-phased trio. I converted seed genotypes to seed

haplotypes using the trio information while keeping “gap” variants as unphased. I then used local phasing conditioned on the seed haplotype to infer phasing of the gap variants using Beagle. I allowed homozygous variants to assist Beagle in making better predictions from the Hidden Markov Model.

To perform neighborhood correction for a seed haplotype unphased variant, I collected three variants each from both upstream and downstream, which are phased in seed haplotype. Then I checked if there was 100% correlation between the phasing present in the seed haplotype to what is predicted by Beagle. This provides an estimate of how well Beagle could have performed in this “local” region. If there is a 100% match, I consider the variant as conditionally phased. If there is not a 100% match, I disregarded the unphased variant in the final haplotype. I tried other window sizes such as 5 and 10 and found no improvement in accuracy.

Local conditional phasing in human GM12878 cells

I coupled HaploSeq analysis and local conditional phasing to increase resolution in GM12878 cells. Local conditional phasing was performed as described earlier on genotypes that are common between GM12878 (ref. 43) and population samples. In addition, as the seed haplotype is not 100% accurate, I marked the seed haplotype phased variants that did not agree with local phasing. These marked variants were made “unphased” as these could be potential errors from HaploSeq. Hence, apart from using neighborhood correction for deciding whether a gap variant needs to be locally phased (as in the simulation), I also

used this information to mark variants in the seed haplotype that could be potentially erroneous.

Figures and Tables

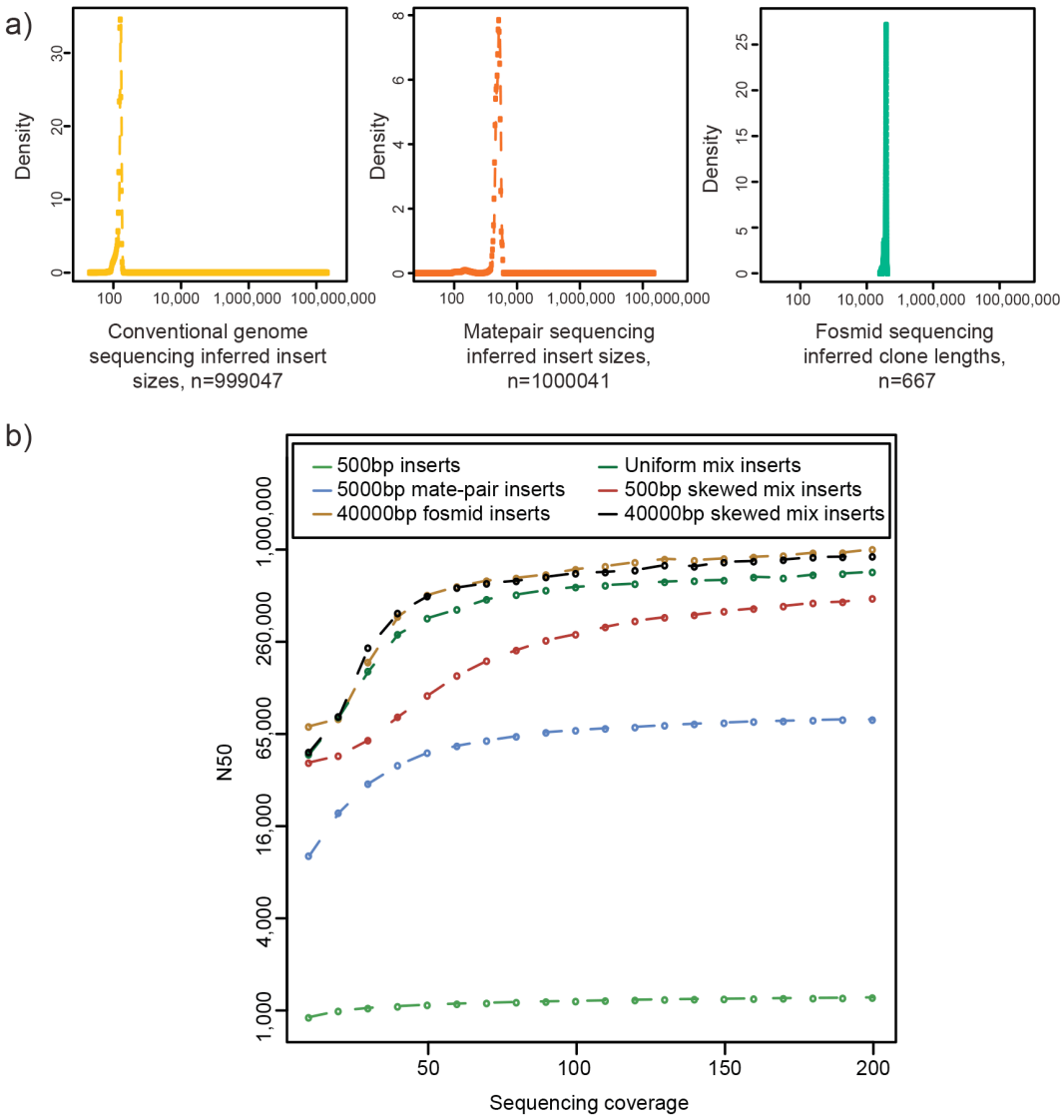

Figure 3-1: The length of haplotype depends on the insert size distributions of the fragments.

a) Inferred insert sizes from conventional genome sequencing (ref. 41), mate-pair (ref. 41) and fosmid clones (ref. 42). The x-axis is in base-pairs (log10 scale). b) Simulations of 100bp paired-end reads at various sequencing coverage for these different datatypes. Each skewed datatype contrains 70% and 10% percent of 40000bp or 500bp, depending on the case. Skew datasets always contain 20% mate-pair. N50 is averaged over 10 simulaitons.

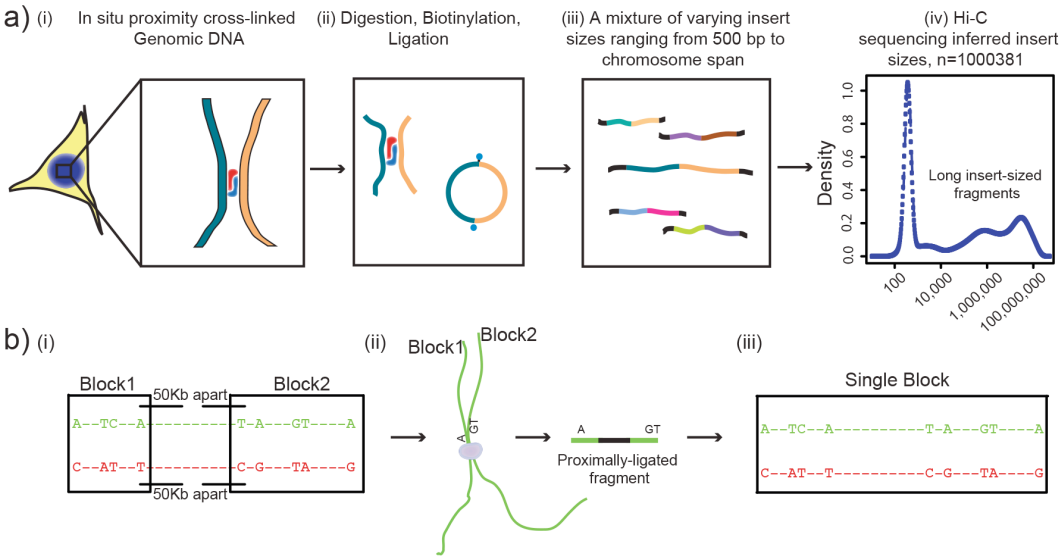

Figure 3-2: Schematic for HaploSeq method for reconstructing haplotypes.

a) Hi-C experiment. In brief, cross-linked chromatin are digested and ligated (i,ii). (iii, iv). Consequently the Hi-C library contains fragments of different insert sizes.The x axis is in base pairs (log10 scale). b) Hi-C reads can build long haplotypes by utilizing combination of small and long insert sized fragments. This cartoon represents a case where two small haplotype blocks can be connected as a single block, as these two are spatial proximal and therefore captured by Hi-C.

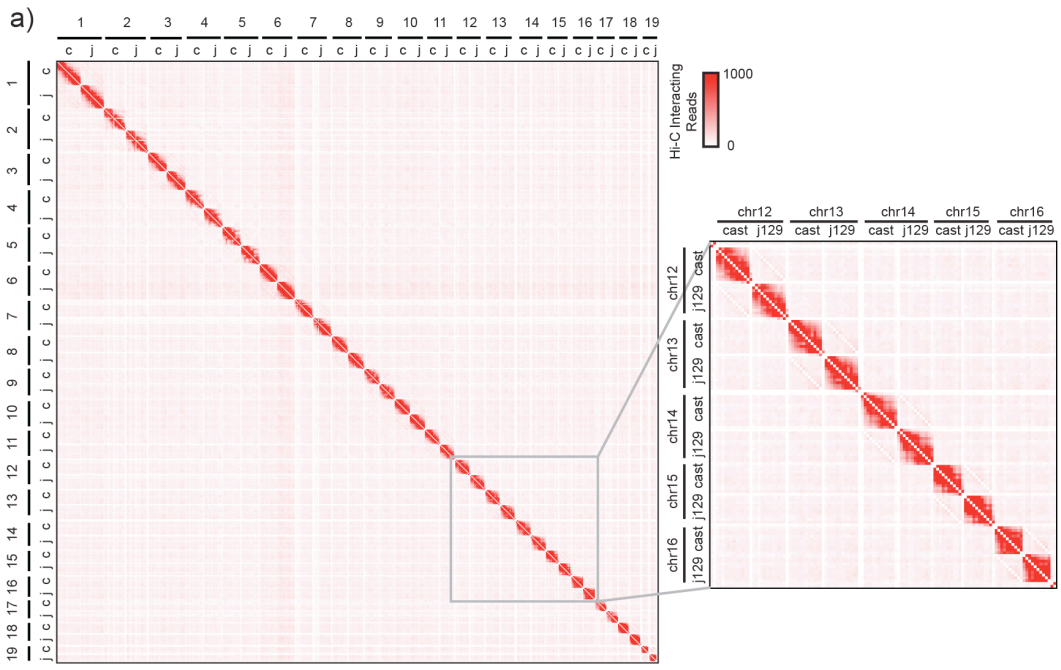

Figure 3-3: Hi-C data demonstrates that the two homologous alleles occupy distinct chromosome territories.

Heat map of whole-genome Hi-C contact frequencies. Hi-C reads originating from the CAST (“c”) or J129 (“j”) genome were distinguished based on the known haplotype structures of the parental strains. The frequency of interactions between each allele of each chromosome was calculated using 10-Mb bin size. The CAST or J129 allele of each chromosome primarily interacts in cis, confirming that the chromosomes territories seen in Hi-C data occur for individual alleles. Inset shows a magnified view of the CAST and J129 alleles for chromosomes 12 through 16.

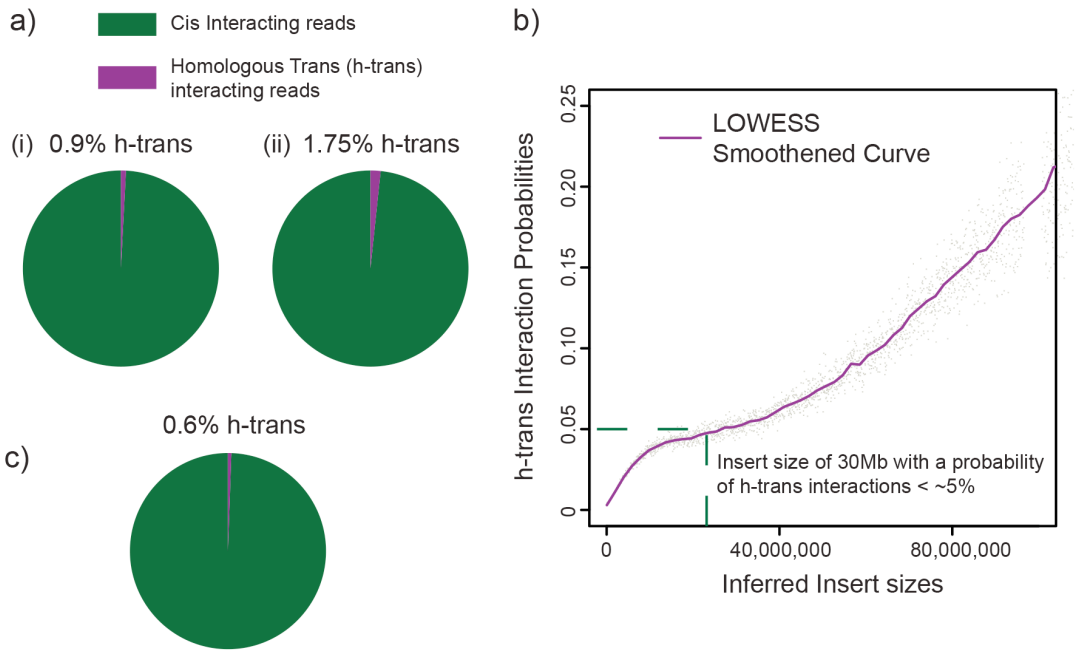

Figure 3-4: Hi-C data is predominantly intrahaplotype.

a) Chart of intrahaplotype (cis) and interhaplotype (h-trans) interaction frequencies. From a priori haplotype information, we distinguish Hi-C read-pairs as interacting in cis (green) and in h-trans (purple). In (i), we used all intrachromosomal reads and in (ii), we excluded all intrachromosomal reads that map with an insert size <1kb, as these are probably short contiguous DNA fragments and are therefore very likely to be in cis. Thus analysis described in (ii) provides a more conservative estimate of h-trans. Comparing these charts, h-trans frequency is at most ~2%. b) Comparison of the h-trans interaction probability as a function of insert sizes. LOWESS fit (purple) was performed at 2% smoothing. c) Similar to b, but excluding reads that have inserts >30 megabases.

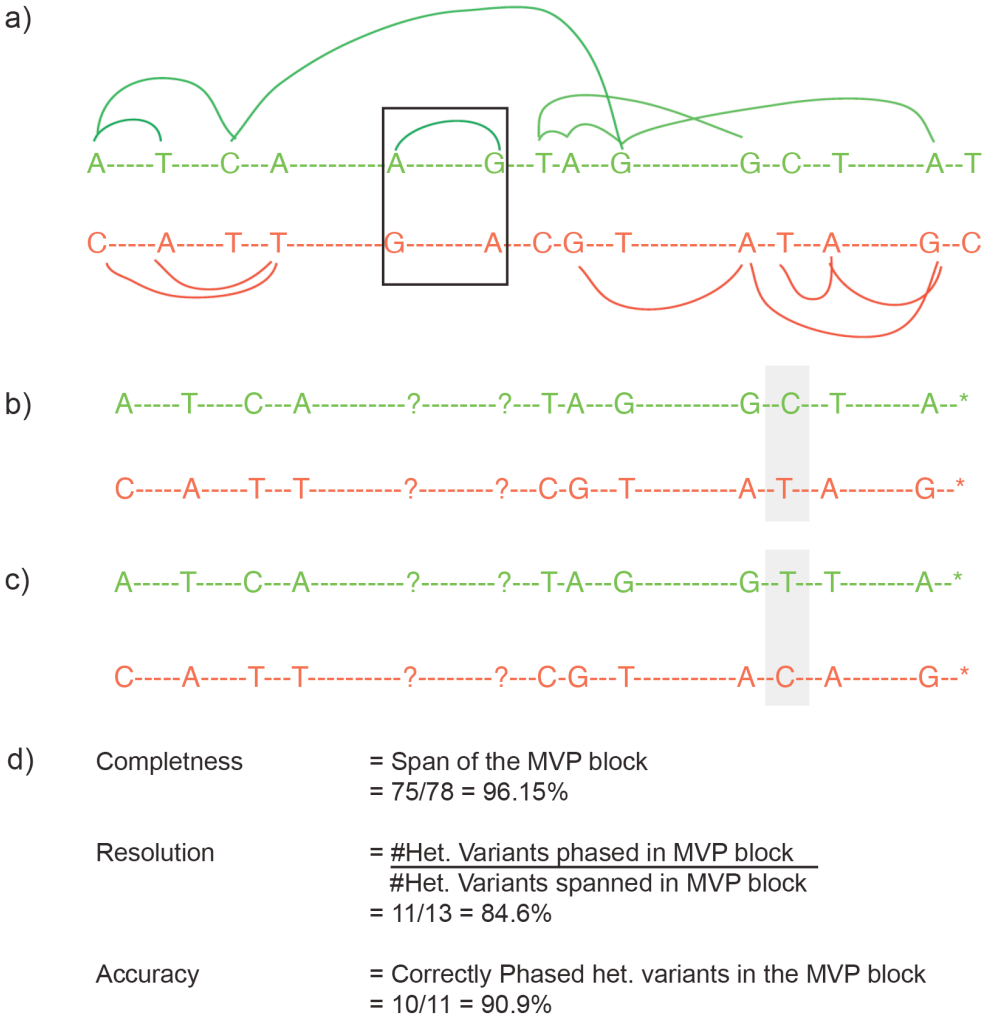

Figure 3-5: Graphical explanation of completeness, accuracy, and resolution in haplotype phasing.

a) Nucletoide bases represent heterozygous SNPs while "-" represents no variability. Considering heterozygous SNPs as nodes, edges are made between nodes that belong to same fragment.This graph system establishes red and green homologous chromosomes (or haplotypes) de-novo. Nevertheless, there can be multiple blocks formed and in this example, we have one large MVP component that spans 96.15% and one other small block that cannot be connected to MVP block (shown in the black edged box). b) Haplotype phasing of the MVP block demonstrating resolution c) True haplotypes known a priori and this knowledge helps to measure the accuracy of predicted de-novo haplotypes. (inaccurate variant phasing is shown at the gray box location) d) Describes the different metrics.

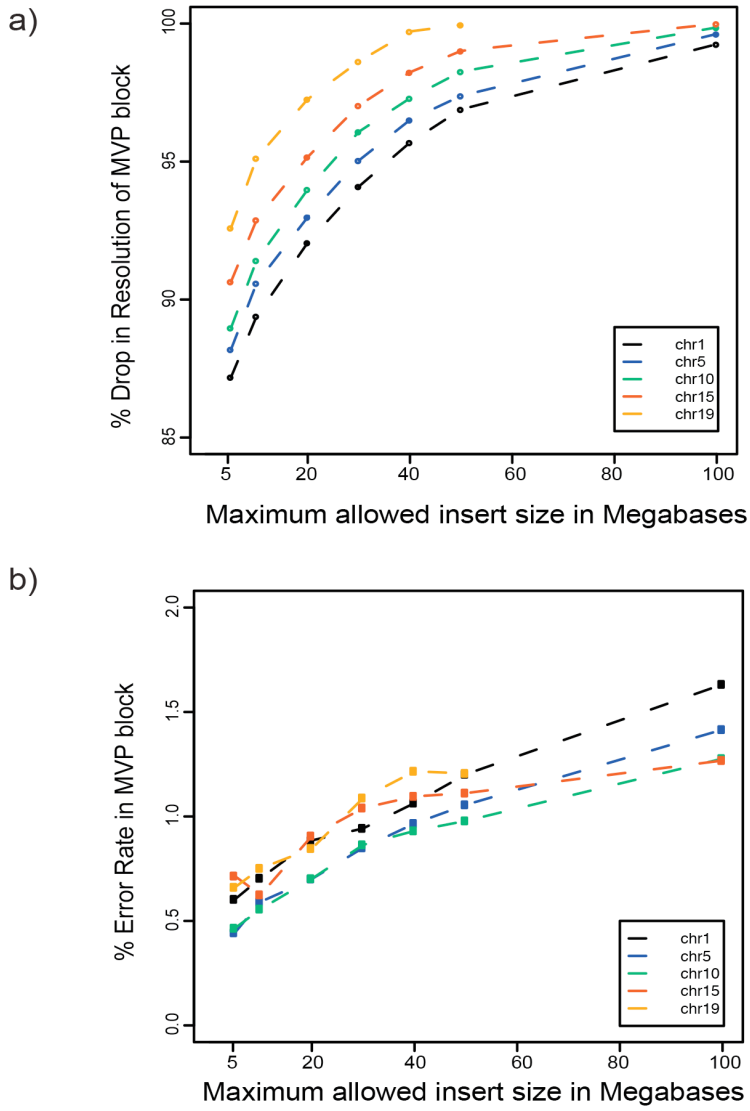

Figure 3-6: Constrained HapCUT model allowing only fragments up to a certain maximum insert size (maxIS).

At higher maxIS, the resolution of MVP block in a) is high but contains lower accuracy in b). Hence, we chose maxIS as 30 megabases to allow acceptable levels of resolution and accuracy. This simulation was performed in different chromosomes in CASTxJ129 system in the low variant density scenario, as this was more close to human applications. This analysis does not incoporate the h-trans probablities, so that the effect of maxIS alone is realized.

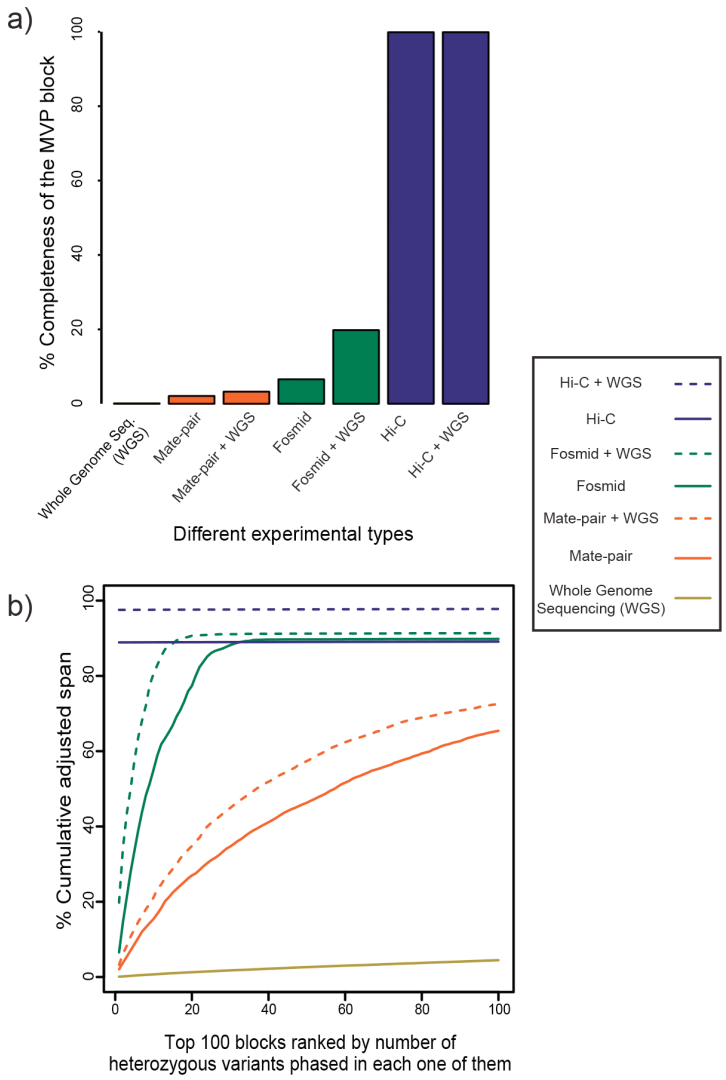

Figure 3-7: HaploSeq resolution can be increased with additional datasets.

a) I simulated 75-bp paired-end sequencing data of conventional shotgun sequencing, mate pair and fosmids at 20× coverage. I subsampled the CAST×J129 data to generate 20× Hi-C fragments.The y axis represents the span of MVP block of chromosome 19. I also combined 20× sequencing coverage for each method with 20× conventional WGS data for a total of 40× coverage to compare methods at a higher coverage. b) Analysis of the adjusted span (AS) of phasing. The AS is defined as the product of span and fraction of heterozygous variants phased in that block. Haplotype blocks were ranked by number of variants phased in each block (x axis is ranking).

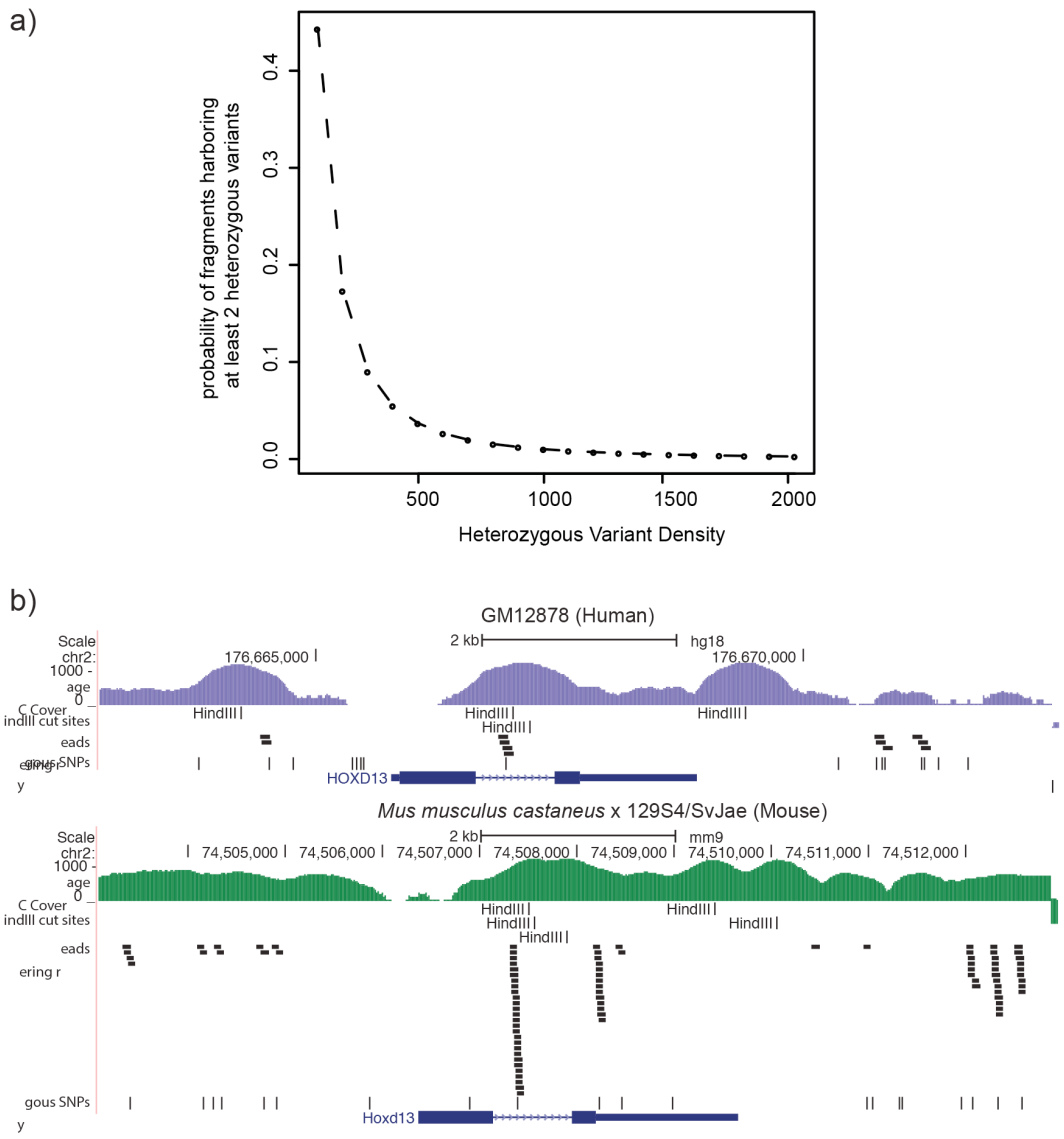

Figure 3-8: Variant density affects the fraction of usable reads and potentially haplotyping.

a) The plot depicts the relationship between variant density and probability of paired-end read pairs harboring at least two heterozygous variants, as only these reads are usefull for phasing. b) The differences in variant frequency between mice (CAST×J129) and humans (GM12878) over the Hoxd13/HOXD13 gene. Also shown in the Hi-C read coverage (log10 scale) over these loci.

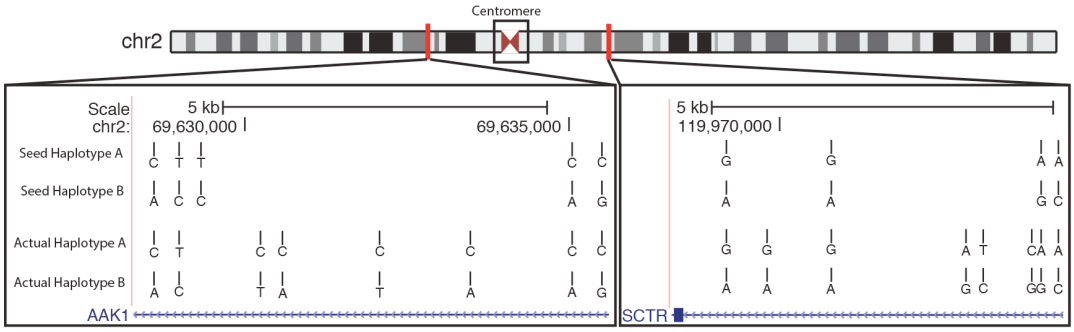

Figure 3-9: HaploSeq generated haplotypes spans accross the centromere.

Hi-C–generated seed haplotypes span the centromere of metacentric chromosomes. Shown are two regions on either side of the centromere of chromosome 2. The two Hi-C generated seed haplotypes are arbitrarily designated as “A” and “B.” The actual haplotypes of the GM12878 individual learned from trio sequencing are shown below designated arbitrarily as “A” and “B.” The Hi-C–generated seed haplotypes match the actual haplotypes on both sides of the centromere. Some variants in the actual haplotype remain unphased, thus contributing to the “gaps” in the seed haplotype. In addition, the actual haplotypes based on trio sequencing may not contain all of the variants from (ref: 43) phased. Therefore, the seed haplotype contains some phased variants not in the trio-phased haplotype (see the third variant in the AAK1 region for example).

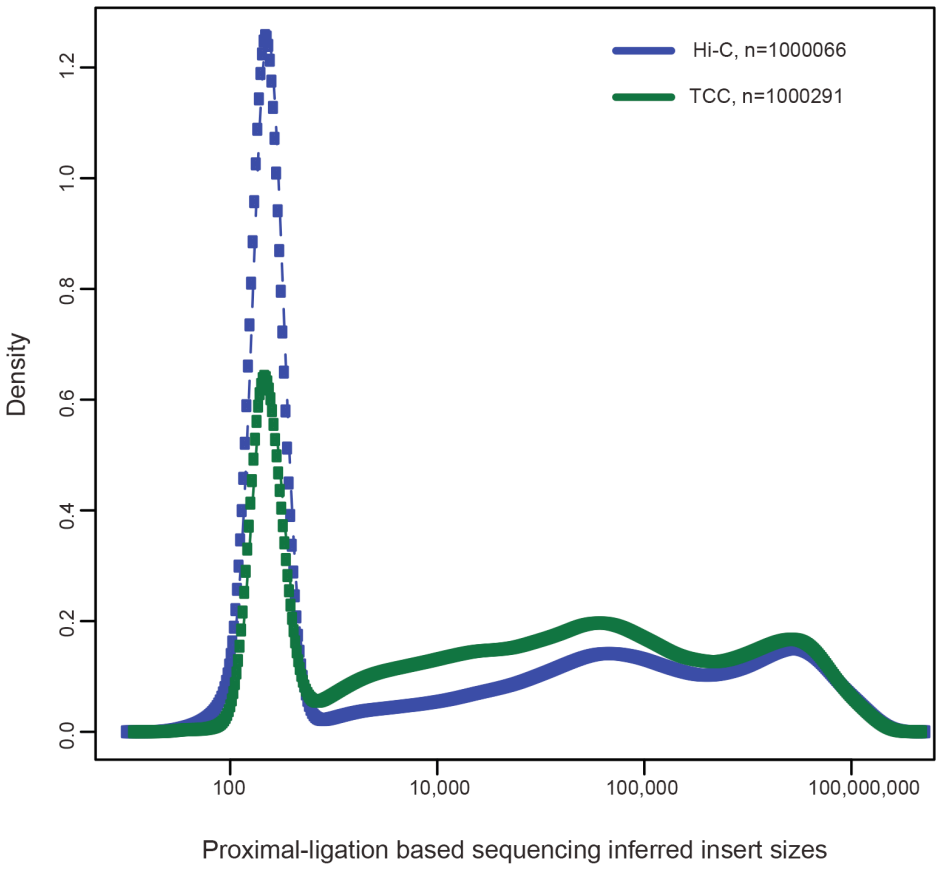

Figure 3-10: Insert size distributions from Hi-C and TCC.

The insert size distributions (log10 scale) from Hi-C and TCC (both taken from ref. 35). TCC has an additional step where ligations are tethered to a solid surface, which are then preferentially captured. Hence, TCC offers more chances to capture true long-range interactions in TCC than in Hi-C experiment. Plots made using random subset of datapoints from chromosomes 1-22.

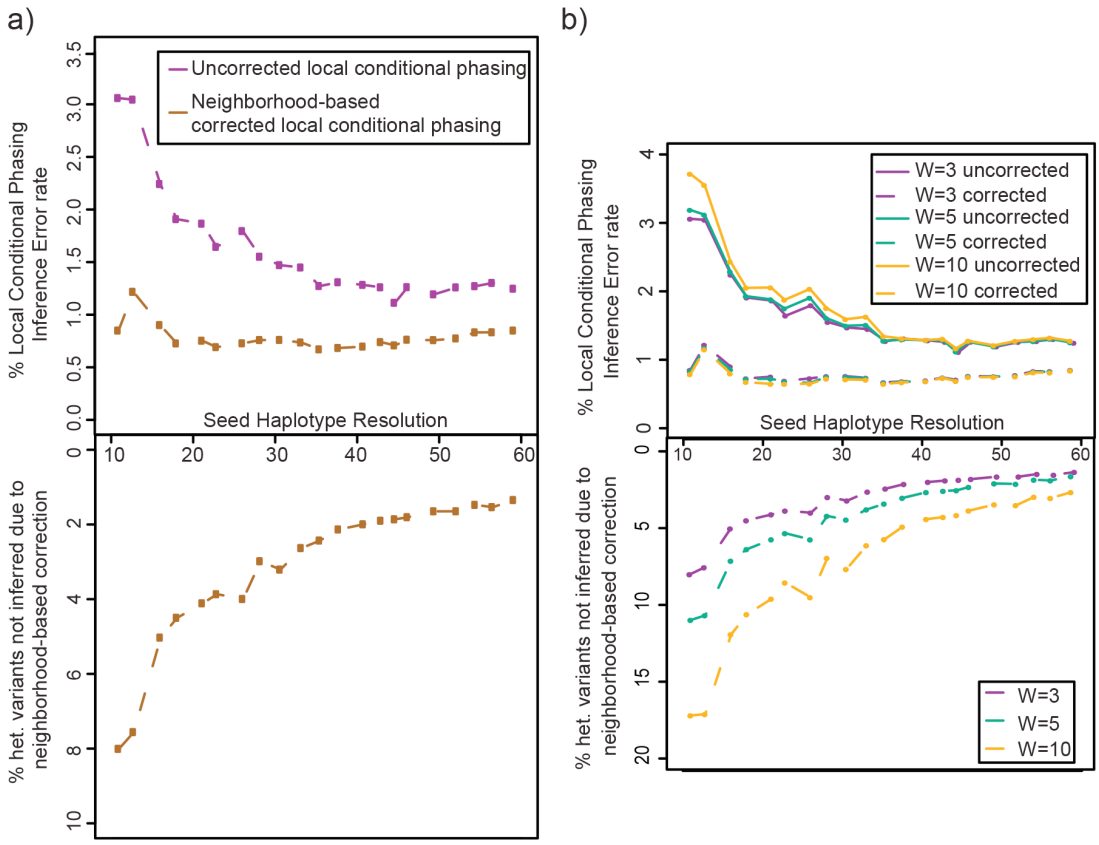

Figure 3-11: Local conditional phasing in human GM12878 cells.

a) The x axis is the chromosome span seed haplotypes resolution generated by simulation. The top panel shows the error rates of local conditional phasing. The bottom panel shows the percentage of variants that remain unphased due to neighborhood correction as a function of resolution. All simulations are done using GM12878 chromosome 1.b) Plots of the affect of window sizes on accuracy and resolution of local condiditonal phasing after neighborhood-window correction. We used window sizes of 3, 5 and 10.

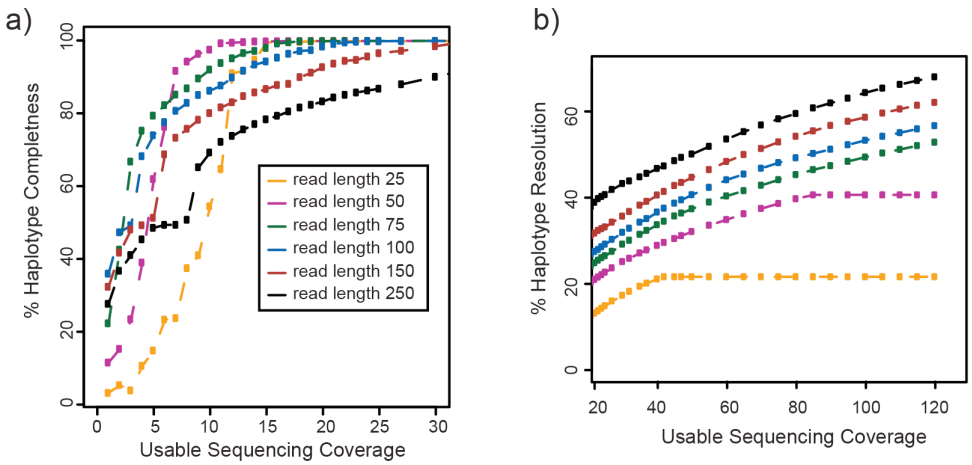

Figure 3-12: Sequencing requiremements for obtaining haplotypes by HaploSeq.

a) Chromosome-spanning seed haplotype (MVP block) at varying parameters of read length and coverage. b) Different combinations of read length and coverage generate high-resolution seed haplotypes. Resolution metric depends on percentage of completeness. For example, for 250 bp reads at 30× coverage, resolution is 45% of the 90% variants spanned in the haplotype. All simulations are done in GM12878 chromosome 1.

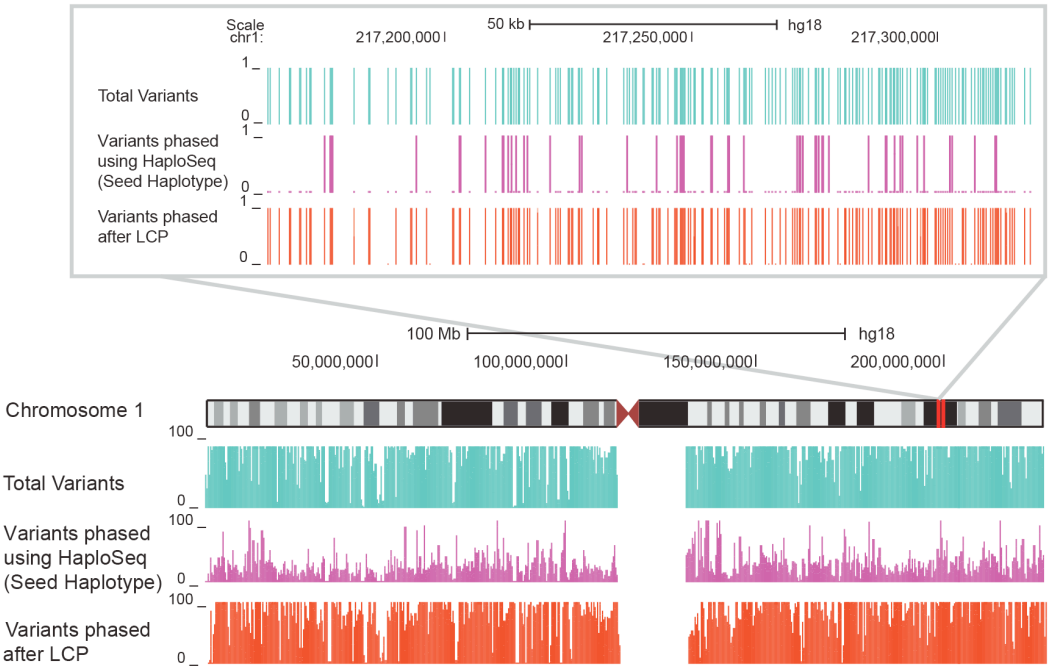

Figure 3-13: HaploSeq coupled with Local conditional phasing (LCP) generates high resolution haplotypes.

UCSC Genome browser shot illustrating all variants (green track), phased variants by HaploSeq (purple track), and phased variants by combining LCP with HaploSeq (red track) in chromosome 1. The track displays the number of heterozygous variants in each category and demonstrates that only a high fraction of variants are phased only after LCP. Top panel, a zoom-in of the browser, showing a binary value for presence (value 1) and absence of a variant (value 0) in that category. A value of 0 in the phased variant track represents unphased variants or "gaps," whereas a value of 1 represents the group of variants that are part of the MVP block. Most of the gaps from the purple track are phased after LCP, as shown in red track.

Table 3-1: Accurate chromosome-span haplotypes in mouse ES cells.

We used HapCUT to phase CASTxJ129 mouse Hi-C data. For every chromosome, we obtain complete chromosome-scale haplotypes (~>99.9%), as seen in theMVP block. Although 99.6% of SNPs have at least one read covering them, ~5% of SNPs do not have reads that connect to them to the MVP block and therefore cannot be phased with respect to MVP block. Consequently, we obtain a resolution of 95%. The accuracy of the haplotype generated is ~99.5%.

Chr	Phasable Span of Chr	Variants spanned in MVP block	%Chr Spanned in MVP block	%Variants Phased in MVP block	% Accuracy of variants phased in MVP block
chr1	194,188,030	1,409,566	100.000	95.231	99.627
chr2	178,746,638	1,109,866	99.997	93.703	99.569
chr3	156,599,306	1,120,125	100.000	94.911	99.639
chr4	152,628,848	1,030,740	99.997	94.366	99.546
chr5	149,536,169	1,063,616	99.999	94.414	99.521
chr6	146,516,752	1,074,301	100.000	96.086	99.674
chr7	149,523,520	965,142	99.999	94.152	99.427
chr8	128,735,517	939,132	99.948	95.060	99.558
chr9	121,070,077	832,047	99.987	94.547	99.600
chr10	126,991,341	980,549	99.996	95.624	99.735
chr11	118,843,488	861,541	99.996	94.612	99.577
chr12	118,256,511	794,128	100.000	94.588	99.515
chr13	117,284,037	858,859	100.000	95.494	99.679
chr14	122,159,750	823,216	99.998	94.707	99.541
chr15	100,494,041	719,697	100.000	94.811	99.618
chr16	95,301,285	711,670	99.898	95.471	99.668
chr17	92,272,062	616,348	99.999	93.669	99.443
chr18	87,771,251	674,750	99.989	95.631	99.599
chr19	58,256,454	411,457	99.869	95.243	99.662

Table 3-2: Lowering variant density resulted in chromosome-scale and accurate haplotypes, but of low resolution.

Table depicting the completness, resolution and accuracy of haplotype reconstruction using HaploSeq analysis in a low variant density scenario in CASTxJ129 system. Variants were sub-sampled in the CASTxJ129 genome to have a heterozygous variant every 1,500 bases, to mimic human scenario.

Chr	% Chr Spanned in MVP block	% Variants Phased in MVP block	% Accuracy of variants phased in MVP block
chr1	99.932	33.428	99.223
chr2	99.975	30.650	99.298
chr3	99.975	32.522	99.079
chr4	99.913	30.948	98.994
chr5	99.947	30.259	99.310
chr6	99.982	35.529	99.223
chr7	99.909	30.841	99.227
chr8	99.879	32.273	99.234
chr9	99.932	31.676	99.338
chr10	99.997	34.246	99.318
chr11	99.997	30.736	99.334
chr12	99.931	31.056	99.061
chr13	99.988	33.793	99.256
chr14	99.627	31.723	99.106
chr15	99.847	33.108	99.168
chr16	99.483	33.586	99.255
chr17	99.920	31.240	99.213
chr18	99.775	33.775	99.174
chr19	99.285	32.464	99.086

Table 3-3: HaploSeq analysis in human GM12878 cells generate complete but low resolution haplotypes.

Table of results of the HaploSeq based haplotype reconstruction in GM12878 cells using variants identified previously (ref: 43). The results show completeness and resolution. In GM12878 cells, we generated ~17x coverage when compared to ~30x in CASTxJ129 system. Therefore, we observe a lower resolution (22%) when compared to low-density CASTxJ129 (32%).

Chr	% Phasable Span of Chr.	Variants spanned in MVP block	% Chr spanned in MVP block	% Variants phased in MVP block
chr1	247,195,920	161,669	99.911	21.596
chr2	242,747,622	174,845	99.984	22.766
chr3	199,384,702	144,914	99.986	23.915
chr4	191,260,971	151,304	99.974	24.687
chr5	180,770,319	139,987	99.890	24.037
chr6	170,883,965	146,307	99.924	28.113
chr7	158,765,244	123,880	99.992	22.819
chr8	146,268,969	115,878	99.912	24.457
chr9	140,252,520	95,981	99.936	22.347
chr10	135,321,315	108,910	99.976	22.544
chr11	134,358,758	104,211	99.984	24.144
chr12	132,273,383	99,405	99.997	21.511
chr13	96,209,726	76,991	99.495	24.260
chr14	88,283,606	68,949	99.973	21.751
chr15	82,077,797	61,540	99.979	22.164
chr16	88,818,477	71,478	99.847	21.507
chr17	78,612,598	54,660	99.745	18.862
chr18	76,114,907	61,146	99.956	22.791
chr19	63,773,223	50,151	99.726	16.706
chr20	62,424,237	49,535	99.745	22.572
chr21	37,193,100	31,891	99.822	22.223
chr22	35,158,263	32,300	99.929	16.464
chrX	151,825,709	64,769	99.982	15.765

Table 3-4: By coupling HaploSeq and local conditional phasing (LCP), we obtain high resolution and accurate haplotypes for GM12878 cells.

While HaploSeq analysis by itself generates low resolution (22%) haplotypes, combining it with LCP enhances resolution to 81%. The second column depicts the enhanced resolution. Owing to strict neighborhood matching during LCP, fraction of resolution is lost (third column). The final column depicts the accuracy of haplotypes.

Chr	% Enhanced MVP Block Res.	% NC based loss in Res.	% Accuracy of variants phased in MVP block
chr1	81.429	2.867	98.164
chr2	81.876	2.224	98.214
chr3	83.665	1.958	98.616
chr4	82.259	1.851	98.459
chr5	82.753	2.498	98.518
chr6	83.308	1.923	98.132
chr7	80.485	2.556	98.445
chr8	84.065	1.643	98.766
chr9	80.058	2.754	98.099
chr10	84.982	1.470	98.743
chr11	84.318	2.597	98.474
chr12	83.593	2.212	98.602
chr13	85.626	1.716	98.429
chr14	82.021	2.121	98.714
chr15	79.897	2.567	98.052
chr16	78.713	2.977	97.945
chr17	75.566	6.591	95.368
chr18	82.409	2.466	98.548
chr19	76.806	5.839	95.985
chr20	83.275	3.414	96.901
chr21	82.657	2.550	98.345
chr22	76.114	6.561	97.843
chrX	72.419	5.981	96.489

Acknowledgements

Chapter 3, in full, is a reprint of the material published in Nature Biotechnology 2013. Siddarth Selvaraj*, Jesse R Dixon*, Vikas Bansal, Bing Ren. “Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing”, Nature Biotechnology 31 (12), 1111-1118. 2013. The dissertation author was a primary investigator and author of this paper. Specifically, the dissertation author performed all of the computational analyses described in the paper.

References

1. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X.Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A. & Rothberg, J.M. The complete genome of an individual by massively parallel DNA sequencing. Nature 452, 872-876 (2008).

2. Pushkarev, D., Neff, N.F. & Quake, S.R. Single-molecule sequencing of an individual human genome. Nature biotechnology 27, 847-850 (2009).

3. Kitzman, J.O., Snyder, M.W., Ventura, M., Lewis, A.P., Qiu, R., Simmons, L.E., Gammill, H.S., Rubens, C.E., Santillan, D.A., Murray, J.C., Tabor, H.K., Bamshad, M.J., Eichler, E.E. & Shendure, J. Noninvasive whole-genome sequencing of a human fetus. Science translational medicine 4, 137ra176 (2012).

4. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., Lin, Y., MacDonald, J.R., Pang, A.W., Shago, M., Stockwell, T.B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S.A., Busam, D.A., Beeson, K.Y., McIntosh, T.C., Remington, K.A., Abril, J.F., Gill, J., Borman, J., Rogers, Y.H., Frazier, M.E., Scherer, S.W., Strausberg, R.L. & Venter, J.C. The diploid genome sequence of an individual human. PLoS biology 5, e254 (2007).

5. Crawford, D.C. & Nickerson, D.A. Definition and clinical importance of haplotypes. Annual review of medicine 56, 303-320 (2005).

6. Petersdorf, E.W., Malkki, M., Gooley, T.A., Martin, P.J. & Guo, Z. MHC haplotype matching for unrelated hematopoietic cell transplantation. PLoS medicine 4, e8 (2007).

7. Studies, N.-N.W.G.o.R.i.A., Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G., Hirschhorn, J.N., Abecasis, G., Altshuler, D., Bailey-Wilson, J.E., Brooks, L.D., Cardon, L.R., Daly, M., Donnelly, P., Fraumeni, J.F., Jr., Freimer, N.B., Gerhard, D.S., Gunter, C., Guttmacher, A.E., Guyer, M.S., Harris, E.L., Hoh, J., Hoover, R., Kong, C.A., Merikangas, K.R., Morton, C.C., Palmer, L.J., Phimister, E.G., Rice, J.P., Roberts, J., Rotimi, C., Tucker, M.A., Vogan, K.J., Wacholder, S., Wijsman, E.M., Winn, D.M. & Collins, F.S. Replicating genotype-phenotype associations. Nature 447, 655-660 (2007).

8. Cirulli, E.T. & Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature reviews. Genetics 11, 415-425 (2010).
9. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., Shendure, J. & Bamshad, M.J. Exome sequencing identifies the cause of a mendelian disorder. Nature genetics 42, 30-35 (2010).

10. Musone, S.L., Taylor, K.E., Lu, T.T., Nititham, J., Ferreira, R.C., Ortmann, W., Shifrin, N., Petri, M.A., Kamboh, M.I., Manzi, S., Seldin, M.F., Gregersen, P.K., Behrens, T.W., Ma, A., Kwok, P.Y. & Criswell, L.A. Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus. Nature genetics 40, 1062-1064 (2008).

11. International Consortium for Systemic Lupus Erythematosus, G., Harley, J.B., Alarcon-Riquelme, M.E., Criswell, L.A., Jacob, C.O., Kimberly, R.P., Moser, K.L., Tsao, B.P., Vyse, T.J., Langefeld, C.D., Nath, S.K., Guthridge, J.M., Cobb, B.L., Mirel, D.B., Marion, M.C., Williams, A.H., Divers, J., Wang, W., Frank, S.G., Namjou, B., Gabriel, S.B., Lee, A.T., Gregersen, P.K., Behrens, T.W., Taylor, K.E., Fernando, M., Zidovetzki, R., Gaffney, P.M., Edberg, J.C., Rioux, J.D., Ojwang, J.O., James, J.A., Merrill, J.T., Gilkeson, G.S., Seldin, M.F., Yin, H., Baechler, E.C., Li, Q.Z., Wakeland, E.K., Bruner, G.R., Kaufman, K.M. & Kelly, J.A. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. Nature genetics 40, 204-210 (2008).

12. Zschocke, J. Dominant versus recessive: molecular mechanisms in metabolic disease. Journal of inherited metabolic disease 31, 599-618 (2008).

13. International HapMap, C., Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., Pasternak, S., Wheeler, D.A., Willis, T.D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S.B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R.C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M.M., Tsui, S.K., Xue, H., Wong, J.T., Galver, L.M., Fan, J.B., Gunderson, K., Murray, S.S., Oliphant, A.R., Chee, M.S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.F., Phillips, M.S., Roumy, S., Sallee, C., Verner, A., Hudson, T.J., Kwok, P.Y., Cai, D., Koboldt, D.C., Miller, R.D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.C., Mak, W., Song, Y.Q., Tam, P.K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C.P., Delgado, M., Dermitzakis, E.T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B.E., Whittaker, P., Bentley, D.R., Daly, M.J., de Bakker, P.I., Barrett, J., Chretien, Y.R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D.J., Sabeti, P., Saxena, R., Schaffner, S.F., Sham, P.C., Varilly, P., Altshuler, D., Stein, L.D., Krishnan, L., Smith, A.V., Tello-Ruiz, M.K., Thorisson, G.A., Chakravarti, A., Chen, P.E., Cutler, D.J., Kashuk, C.S., Lin, S., Abecasis, G.R., Guan, W., Li, Y., Munro, H.M., Qin, Z.S., Thomas, D.J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L.R., Clarke, G., Evans, D.M., Morris, A.P., Weir, B.S., Tsunoda, T., Mullikin, J.C., Sherry, S.T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D.R., Suda, E., Rotimi, C.N., Adebamowo, C.A., Ajayi, I., Aniagwu, T., Marshall, P.A., Nkwodimmah, C., Royal, C.D., Leppert, M.F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I.F., Knoppers, B.M., Foster, M.W., Clayton, E.W., Watkin, J., Gibbs, R.A., Belmont, J.W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G.M., Wheeler, D.A., Yakub, I., Gabriel, S.B., Onofrio, R.C., Richter, D.J., Ziaugra, L., Birren, B.W., Daly, M.J., Altshuler, D., Wilson, R.K., Fulton, L.L., Rogers, J., Burton, J., Carter, N.P., Clee, C.M., Griffiths, M., Jones, M.C., McLay, K., Plumb, R.W., Ross, M.T., Sims, S.K., Willey, D.L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J.C., L'Archeveque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A.L., Brooks, L.D., McEwen, J.E., Guyer, M.S., Wang, V.O., Peterson, J.L., Shi, M., Spiegel, J., Sung, L.M., Zacharia, L.F., Collins, F.S., Kennedy, K., Jamieson, R. & Stewart, J. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851-861 (2007).

14. Genomes Project, C., Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. & McVean, G.A. A map of human

genome variation from population-scale sequencing. Nature 467, 1061-1073 (2010).

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

16. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prufer, K., de Filippo, C., Sudmant, P.H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R.E., Bryc, K., Briggs, A.W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M.F., Shunkov, M.V., Derevianko, A.P., Patterson, N., Andres, A.M., Eichler, E.E., Slatkin, M., Reich, D., Kelso, J. & Paabo, S. A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222-226 (2012).

17. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R. & Chess, A. Widespread monoallelic expression on human autosomes. Science 318, 1136-1140 (2007).
18. Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K.T., Jonasdottir, A., Frigge, M.L., Gylfason, A., Olason, P.I., Gudjonsson, S.A., Sverrisson, S., Stacey, S.N., Sigurgeirsson, B., Benediktsdottir, K.R., Sigurdsson, H., Jonsson, T., Benediktsson, R., Olafsson, J.H., Johannsson, O.T., Hreidarsson, A.B., Sigurdsson, G., Consortium, D., Ferguson-Smith, A.C., Gudbjartsson, D.F., Thorsteinsdottir, U. & Stefansson, K. Parental origin of sequence variants associated with complex diseases. Nature 462, 868-874 (2009).

19. Xie, W., Barr, C.L., Kim, A., Yue, F., Lee, A.Y., Eubanks, J., Dempster, E.L. & Ren, B. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell 148, 816-831 (2012).

20. McDaniell, R., Lee, B.K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., Keefe, D., Collins, F.S., Willard, H.F., Lieb, J.D., Furey, T.S., Crawford, G.E., Iyer, V.R. & Birney, E. Heritable individual-specific and allele-specific chromatin signatures in humans. Science 328, 235-239 (2010).

21. Fan, H.C., Wang, J., Potanina, A. & Quake, S.R. Whole-genome molecular haplotyping of single cells. Nature biotechnology 29, 51-57 (2011).

22. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American journal of human genetics 81, 1084-1097 (2007).

23. Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., Haas, J., Robasky, K., Zaranek, A.W., Lee, J.H., Ball, M.P., Peterson, J.E., Perazich, H., Yeung, G., Liu, J., Chen, L., Kennemer, M.I., Pothuraju, K., Konvicka, K., Tsoupko-Sitnikov, M., Pant, K.P., Ebert, J.C., Nilsen, G.B., Baccash, J., Halpern, A.L., Church, G.M. & Drmanac, R. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487, 190-195 (2012).

24. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics 24, i153-159 (2008).

25. Kitzman, J.O., Mackenzie, A.P., Adey, A., Hiatt, J.B., Patwardhan, R.P., Sudmant, P.H., Ng, S.B., Alkan, C., Qiu, R., Eichler, E.E. & Shendure, J. Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nature biotechnology 29, 59-63 (2011).

26. Suk, E.K., McEwen, G.K., Duitama, J., Nowick, K., Schulz, S., Palczewski, S., Schreiber, S., Holloway, D.T., McLaughlin, S., Peckham, H., Lee, C., Huebsch, T. & Hoehe, M.R. A comprehensively molecular haplotype-resolved genome of a European individual. Genome research 21, 1672-1685 (2011).

27. Duitama, J., McEwen, G.K., Huebsch, T., Palczewski, S., Schulz, S., Verstrepen, K., Suk, E.K. & Hoehe, M.R. Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. Nucleic acids research 40, 2041-2053 (2012).

28. Kaper, F., Swamy, S., Klotzle, B., Munchel, S., Cottrell, J., Bibikova, M., Chuang, H.Y., Kruglyak, S., Ronaghi, M., Eberle, M.A. & Fan, J.B. Whole-genome haplotyping by dilution, amplification, and sequencing. Proceedings of the National Academy of Sciences of the United States of America 110, 5552-5557 (2013).

29. Yang, H., Chen, X. & Wong, W.H. Completely phased genome sequencing through chromosome sorting. Proceedings of the National Academy of Sciences of the United States of America 108, 12-17 (2011).

30. Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K. & Song, Q. Direct determination of molecular haplotypes by chromosome microdissection. Nature methods 7, 299-301 (2010).

31. Kirkness, E.F., Grindberg, R.V., Yee-Greenbaum, J., Marshall, C.R., Scherer, S.W., Lasken, R.S. & Venter, J.C. Sequencing of isolated sperm cells for direct haplotyping of a human genome. Genome research 23, 826-832 (2013).

32. Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J. & Schork, N.J. The importance of phase information for human genomics. Nature reviews. Genetics 12, 215-223 (2011).

33. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. Science 295, 1306-1311 (2002).

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

35. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nature biotechnology 30, 90-98 (2012).
36. Krueger, C., King, M.R., Krueger, F., Branco, M.R., Osborne, C.S., Niakan, K.K., Higgins, M.J. & Reik, W. Pairing of homologous regions in the mouse genome is associated with transcription but not imprinting status. PloS one 7, e38983 (2012).

37. Browning, B.L. & Browning, S.R. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. Genetics 194, 459-471 (2013).

38. He, X., Fuller, C.K., Song, Y., Meng, Q., Zhang, B., Yang, X. & Li, H. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. American journal of human genetics 92, 667-680 (2013).

39. Zeng, D. & Lin, D.Y. Estimating haplotype-disease associations with pooled genotype data. Genetic epidemiology 28, 70-82 (2005).

40. Chapman, J.M., Cooper, J.D., Todd, J.A. & Clayton, D.G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Human heredity 56, 18-31 (2003).

41. Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S. & Jaffe, D.B. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America 108, 1513-1518 (2011).

42. Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N.A., Tsang, P., Newman, T.L., Tuzun, E., Cheng, Z., Ebling, H.M., Tusneem, N., David, R., Gillett, W., Phelps, K.A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J.D., Korn, J.M., McCarroll, S.A., Altshuler, D.A., Peiffer, D.A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D.A., Mullikin, J.C., Wilson, R.K., Bruhn, L., Olson, M.V., Kaul, R., Smith, D.R. & Eichler, E.E. Mapping and sequencing of structural variation from eight human genomes. Nature 453, 56-64 (2008).

43. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D. & Daly, M.J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 43, 491-498 (2011).

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

45. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M.A. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303 (2010).

Chapter 4: Analysis of haplotype-resolved gene regulation patterns in human ES cells and ES-derived cell-types

Abstract

Recent collaborative projects such as the ENCODE and Roadmap Epigenome have allowed annotation of regulatory elements and subsequent investigation of their role in cellular differentiation and lineage specification. However, these analyses are limited in two aspects. First, the functional maps contain mixture information of the two haploids and thus epigenetic and genetic differences between the haplotypes are ignored. Second, current analyses are limited in recognizing the role of distal gene regulation. To address these challenges, we performed Hi-C in H1 human embryonic stem cells and 4 H1-derived cells from diverse developmental lineages, as it can inform both haplotype and 3D genome patterns. We integrated previously obtained maps of chromatin accessibility, DNA methylation, histone modifications, and gene expression to delineate aspects of gene regulation in an allelic context. By phasing over 93.5% of alleles, the haplotype-resolved genome revealed widespread allelic gene expression patterns. In addition, we observe a strong correlation among allelic transcription and allelic chromatin states of promoters and distal acting enhancers. By correlating allelic regulatory states and allelic gene activity, our study demonstrates new insights on combinatorial functional interactions of gene regulation.

Introduction

Human cellular differentiation is a complex process that harbors unique gene expression patterns in each cell type[1-3]. It is increasingly being accepted that cell type specific gene regulation patterns are facilitated by dynamic changes in epigenome[4-9]. For example, DNA Methylation at promoters has been shown to inhibit expression of lineage-specific genes and regulate imprinting regions[10-12]. On the same lines, other epigenomic aspects such as histone modifications have also suggested to play a critical role in animal development[13]. For example, mice with depleted histone acetyltransferase p300/CBP are lethal[14].

To systematically study the role of epigenetic mechanisms in human development, the Roadmap Epigenome project profiled DNA Methylation, core histone marks, chromatin accessibility and gene expression in H1 human embryonic stem cells (hESCs) and four-hESC derived lineages[4]. In particular, Mesendoderm, Mesenchymal Stem Cells, Neural Progenitor Cells, and Trophoblast were chosen as they represent extra-embryonic and embryonic lineages, including cells at early and late stages of development. Utilizing these datasets, lineage specific regulatory elements were defined using which distinct epigenetic mechanisms for regulation of early and late differentiated stages were reported, clearly showing crucial role of epigenetics in human cellular differentiation[4, 6].

Along with epigenomes, several groups have demonstrated the role of 3D genome structure in regulating cell-type specific gene expression[15,16]. For example, it has been shown that 3D genome can facilitate chromatin interactions

among distal regulatory elements such as enhancers and target genes[17,18]. However, the vast majority of studies that analyze gene regulation pattern have not performed integrative analyses of epigenome and 3D genome. In addition, these analyses could be confounded by the fact that each dataset contains mixture information of the two haplotypes. Specifically, current studies are limited in reporting allele specific regulatory events and allelic gene expression. For example, imprinting genes are known to express in an allelic fashion[10], however the scope of such allele-specific genes are poorly understood in the context of cellular differentiation.

As Hi-C can inform both 3D structure[19-21] and haplotypes[22], we have currently performed Hi-C in each of these 5 lineages to integrate analyses of chromatin structure, epigenome and gene expression in a haplotype resolved context. By analyzing allele-resolved gene expression patterns, we identify widespread allelic biases in gene expression in each lineage, consistent with recent reports in individual cell types. In total, 24% of genes in the genome for which we can reliably detect allele-resolved expression show an allelic bias, indicating that this phenomenon is pervasive throughout the genome. Allele biased patterns of gene expression are well correlated with allelic biased chromatin state at distal acting enhancer elements and long-range chromatin interactions between these elements and the target genes. Our results demonstrate a strong relationship between dynamic chromatin architecture and dynamic chromatin states, together coordinating gene regulation in an allelic context. Taken together, our study shows a combinatorial functional interaction

between regulatory elements and genes, facilitated by 3D genome and haplotype analyses.

Results

Generating complete haplotype structures for H1 cell line

We performed Hi-C[20] experiments in H1 hESCs and each of the four H1-derived lineages. We obtained a total of 3.85 billion unique read pairs, with on average 770 million unique read pairs split between two biological replicates for each cell type (Table 4-1). Using HaploSeq[22], I generated chromosome span haplotypes for H1 by combining the Hi-C datasets across all of the H1-lineages, and whole genome sequencing to maximize phasing resolution (Fig. 4-1a). In total, I was able to generate haplotypes incorporating ~93.5% of all heterozygous variants in the H1 genome. To evaluate the accuracy of the haplotype predictions, I performed HaploSeq using reads from Hi-C alone and checked its concordance with independent datasets such as whole genome sequencing and mRNA-Seq (Fig. 4-1b). As, the concordance rates for the H1 genome are similar to the error rates we found in previous work using cell lines where haplotypes were known a priori where the accuracy of phasing could be calculated explicitly, we believe that the haplotypes predictions of H1 genome are of high quality[22].

Having obtained complete, accurate, and high-resolution haplotypes, we analyzed various genome wide datasets in an allele resolved context[4]. We re-aligned mRNA-sequencing, ChIP-sequencing for histone modifications, MethylC-Sequencing, and DNaseI hypersensitivity sequencing datasets for each of the

H1-derived lineages and determined which reads arose from which haplotype (Methods). Of note, as we have only haplotype information for the H1 individual, we cannot determine which allele is the maternal or paternal copy. Therefore, we arbitrarily defined the two parental haplotypes for each chromosome to be from the “p1” allele and “p2” allele. As another metric to check the accuracy of haplotypes, I checked if mechanisms behind an imprinting region could be recapitulated using various epigenetic datasets. Indeed SNRPN, a known DNA Methylation based imprinted gene cluster[23], is expressed only in p1 allele as supported by active H3K4me3 histone mark in p1 and inactive methylated promoter at p2 allele (Fig. 4-2). These datasets therefore allow for the systematic determination of variability in gene expression and chromatin state of cis-regulatory elements between alleles.

Identifying allelic events

As allelic events could be a result of biased mapping strategies, we followed a multi-step process to accurately identify allelic events. Besides, mapping each of the datasets to a heterozygous variant masked human reference genome, I simulated reads spanning each of the variant to estimate mapping biases. SNPs and Indels that showed >5% and >10% biases respectively, were excluded from all downstream analyses, as these variants potentially show an inherent mapping bias. Second, variants that demonstrated >3 standard deviations or significant binomial variation (FDR 5%) of genome sequencing coverage above the mean haplotype coverage, were removed as

potential sources of copy number variation. Next, we excluded any heterozygous variant with a genotype p-value greater than 0.05 after Benjamini correction, as these can be inherently homozygous in nature (Methods). Using the final list of heterozygous variants, we employed different statistical methods such as negative binomial (allelic genes), binomial (allelic chromatin states), and hyper-geometric test (DNA Methylation) to evaluate the allelic status (Methods).

Widespread allelic imbalances in gene expression

Previous studies of allele-resolved gene expression have identified allelic imbalances in expression of a given gene between two alleles[24, 25]. However, most previous studies of allele resolved gene expression focus on only a limited number of cell types, most often lymphoblastoid cell lines. Therefore, it remains unclear the degree to which allele-biased gene expression varies among different lineages of a single individual. To address this issue, I identified allelic biases in gene expression across the five H1 lineages examined in this study. I identified a total of 1,787 genes that showed allelic bias in gene expression between the two alleles in any cell type (FDR 10%, Fig. 4-3a). As only genes that contain exonic SNPs and can possibility be analyzed for allele specific expression, this actually represents 24% of all genes for which we can detect allelic expression (Fig. 4-3a). This suggests that allele biased gene expression is pervasive throughout the H1 human genome. In addition, most of the allelic differences in expression were less than 4-fold (Fig. 4-3b), indicating that the majority of allelic differences in expression were not “on/off” events, but instead reflecting changes in the relative

level of expression from each allele. By performing k-means clustering on the patterns of expression of allelic biased genes across cell types, we observed that genes that show bias in expression contain both lineage specific and constitutively expressed genes (Fig. 4-3c). However, allele biased genes do not appear to be enriched among annotated lists of either housekeeping or lineage-restricted genes as compared with non-allele biased genes (Fig. 4-3d).

We were also interested in characterizing if the patterns of bias between the two alleles vary between cell types. For genes that are expressed exclusively in only one or two lineages, allele bias could only occur in a cell type specific manner. Therefore, we focused our analysis on genes were we could detect expression across all 5 lineages. By performing K-means clustering of the patterns of bias among these constitutively expressed genes, we can observe that some allelic genes show constitutive allelic bias, whereas others show cell-type variable patterns of bias (Fig. 4-3e,f). Cell-type variability in allelic bias appears to largely be related to a gain of allelic bias in a particular lineage.

Allelic bias is enriched among imprinted genes

While imprinted genes are enriched in the set of allelic biased genes, they make up only a small fraction (~1%) of the allelic-biased genes (Fig. 4-4a,b). Further, as imprinted genes are generally regulated in clusters[23], I also assessed whether allele biased genes in general tend to occur in clusters. While allele-biased genes, tend to locate closer to other allele biased genes (Fig. 4-4c, p=0.0482 Wilcox rank sum test), the differences are very subtle, suggesting that

the majority of allele-biased gene expression appears not occur in clusters. Therefore, it appears that that most of the allelic gene expression is due to mechanisms other than genomic imprinting.

Allelic promoter bias correlates with allelic transcription

As cis-regulatory elements such as promoters and enhancers are known to play critical role in gene regulation[26, 27], we hypothesized that allelic gene expression could be at least partially explained by sequence variations in these cis elements. To test this hypothesis, I identified SNPs in the H1 lineage that showed any kind of allele specific bias when considering histone acetylation, histone methylation, or DNase I hypersensitivity. We observe that SNPs that show some kind allelic bias are indeed closer to allele-biased genes than unbiased SNPs (Fig. 4-5a). Encouraged by this result, we characterized DNA methylation or chromatin modification state at the promoters of allele biased genes to check if allelic transcription correlates with chromatin state of promoter (Fig. 4-5b,c). Specifically, only 247 (14%) out of 1,787 allele-biased genes contain allelic biased SNPs in their promoter region at least one lineages and are therefore amenable to this analysis. Of these 247 genes, a majority contains either active or repressive marks at their promoter (Fig. 4-5b), supporting a role for allele specific activation or repression of the promoter in the establishment of the allelic expression status of these genes. The concordance with repressive chromatin states is largely due to the allelic biased DNA methylation patterns. Notably, on average 29% of genes that have allele-biased expression show no

evidence of allelic-bias at their promoter region, despite the presence of SNPs in the promoter with the potential to distinguish allele specific activity, suggesting the use of alternative mechanisms, such as regulation through distal-acting enhancers.

Patterns of allelic enhancer sites

As allele biased expression could be the result of allele-biased events at distal enhancer elements, we analyzed allelic patterns of histone acetylation and DNase I HS (DHS) at previously predicted enhancer elements in the H1 and H1-derived cell lines[4]. We were able to identify 1,589 enhancers that displayed allele-biased chromatin state in at least one of the 5 cell lines analyzed (Fig. 4-6a). Several lines of evidence suggest that these allele-specific enhancers are contributing to gene regulation. First, enhancers that show allelic DHS or acetylation show depleted levels of DNA methylation (Fig. 4-6b). Second, these enhancers are generally located closer to genes that also show allele biased expression when compared with enhancers that lack allele bias (Fig. 4-6c). To systematically analyze allelic enhancers with respect to genes, it is critical to link enhancers to target genes. However, as enhancers can regulate distal and often multiple genes, finding true target genes for enhancers have been challenging.

Using C-based technologies to link allelic enhancers and target genes

We hypothesized that allelic enhancers, which are spatially proximal to allelic target genes, are more likely to be involved in gene regulation. To quantify

spatial proximity, we developed a computational strategy using Hi-C data (Methods). Briefly, we divided the genome into 5kb bins and calculated interaction frequency for every promoter-enhancer pair using normalized Hi-C data. Next, we summed up interaction frequencies at multiple resolutions of enhancers so as to enrich for Hi-C signal. We then used a Weibull distribution to estimate significance values and true enhancer-promoter interactions were chosen based on 0.1% FDR (Methods). To validate predicted enhancer-promoter interactions we compared the interaction frequency scores to the previously published 5C dataset[17]. We observe strong correlative patterns between 5C and our interaction frequency scores (Fig. 4-7a). In addition, we employed high-resolution 4C-seq[28] from 6 allele biased enhancer elements. We developed a distance dependent LOWESS regression model of the quantile normalized 4C-seq interaction frequencies (using 4cseqpipe[28]) in order to identify “specific” interactions between the allele biased enhancers and the surrounding regions (Fig. 4-7b).

Spatially proximal allelic enhancers correlate to transcription

Using the predicted enhancer-promoter interactions, we observed that there is a greater correlation between allelic enhancer state and allelic gene expression when the gene and enhancer are spatially proximal as defined by strong Hi-C interaction scores (Fig. 4-8a). Most of these allelic enhancers are likely regulating genes at long distances. For instance, only 10% of allelic expressed genes have an allelic enhancer within 20kb (Fig. 4-8b). In contrast,

66% of the 640 allelic gene-enhancer pairs analyzed display strong Hi-C interactions with allelic enhancers located greater than 20kb away (Fig. 4-8b). In addition, by considering loci that have 4C-Seq interaction frequencies >2.5x over the LOWESS expected model, we observe 4 out of 6 tested allelic enhancers to be spatially proximal to allelic genes (Fig. 4-8c). While one locus showed interaction frequencies to allelic gene with <2.5fold LOWESS enrichment, other loci showed interactions to MT1H and MT1G genes that was not amenable to allelic analyses (Fig. 4-8c). In summary, we observe specific spatial contacts between enhancers and target genes, indicating that allele biased enhancers likely are regulating allele biased genes; though it remains possible that a minority of allele biased enhancers are not regulating any target genes.

Allelic bias may contribute to human health and disease

To understand associations between allele-biased state and common diseases or phenotypes, I identified all SNPs in the GWAS catalog[29] that were present as heterozygotes in the H1 genome. I expanded the H1 GWAS list by including variants linked in Linkage disequilibrium (r2 > 0.8). Several observations suggest that allelic activity may contribute to phenotypic diversity. For one, GWAS SNPs are closer to allele-biased genes than would be expected at random (Fig. 4-9a). Second, we analyzed the enrichment of active chromatin marks (histone acetylation, DHS, H3K4me1, H3K4me3, H3K36me3) at GWAS SNPs in the H1 genome. Specifically, we compared the enrichment of these marks on the risk versus non-risk allele in H1, and we observe that the risk

alleles have a slightly lower chromatin activity when compared with the non-risk alleles (Fig. 4-9b), suggesting that these variants may be associated with a moderate loss or reduction in activity. For example, one locus identified corresponds to one of the allele biased enhancers we used for 4C-Seq analysis (Fig. 4-9c). In this case, a SNP linked to Systemic Lupus Erythematosis is located within an allele-biased enhancer in an intron of the PXK gene. At this locus, the risk allele shows reduced histone acetylation relative to the non-risk allele. In addition, our 4C-seq analyses indicates that this variant is forms specific interactions with the promoter of the PXK gene which shows allele bias in expression with reduced expression on the same haplotype as the risk allele (Fig. 4-9c), suggesting a potential molecular mechanism for this genetic variant.

Allelic bias occur from both parental haplotypes

As we demonstrate the gene regulation patterns in a haplotype-resolved context, we also wanted to check if there is any bias in allelic bias towards any parent. Although we cannot determine which haplotype is paternal or maternal, we can infer parental biases in allelic events. In particular, we assessed for each chromosome the fraction of allelic bias present on the p1 allele for allele biased genes and allele biased enhancers as called by either allelic DHS or allelic acetylation. Although, there is some degree of variability in bias between the p1 and p2 alleles for each chromosome, none of the chromosomes show a statistically significant bias in allelic events to either allele (Fig. 4-10a,b,c). The greater variability in the allelic acetylation and allelic DHS compared with allelic

genes is likely a product of the fact that there are fewer elements called as allele biased on each chromosome for these relative to allelic genes, and therefore calculating the fraction of elements on a given allele is subject to greater variability (Fig. 4-10b,c). Therefore, our data suggests that allelic activities are contributed from both the parents in a similar proportion.

Discussion

We have presented here Hi-C interaction maps in H1 hESC and four H1-derived lineages. These maps have allowed for comprehensive reconstruction of chromosome-span haplotypes for the H1 genome, enabling analysis of gene expression and chromatin states of regulatory elements. Furthermore, as regulatory elements can be distal to target genes, we have used Hi-C and 4C-Seq interaction maps to link cis-regulatory elements to genes and therefore perform an integrative analysis of genome sequence, structure and epigenome. Analyzing these datasets in a haplotype resolved context have revealed new insights on allelic gene regulation.

We have observed extensive allele specific gene expression. Nearly a quarter of genes appear to have an allelic bias in at least one of the cell lines analyzed. In addition, the transcription of majority of these allelic genes can be linked to allelic chromatin states of cis-regulatory elements, such as promoters and enhancers. We cannot currently determine if allelic activities at these functional sites are due to genetic, epigenetic, or their interplay. Regardless, our

results reveal a coordinated activity among genome sequence and structural features.

Analysis of gene regulation in an allelic context has several implications for our understanding of the mechanisms of human development. For instance, phenomena such as compound heterozygosity were well described for coding variation. Our results suggest that non-coding variation in distal regulatory elements may also contribute to potential instances of compound heterozygosity. This underscores the importance of obtaining long-range haplotype information for an individual in order to understand the consequence of inheriting distal acting variants. Inevitably, these studies will need to become routine in order to understand the effects of distal acting non-coding variation on gene expression.

As the two haplotypes differ primarily in genetic and epigenetic aspects, they can be contrasted with changes in gene expression among a population of individuals to understand the basis of human disease. Such studies can allow build predictive models of gene regulation utilizing aspects of genome structure and function of sequence-based regulatory elements. As allele-biased expression is widespread in the genome of an individual, this suggests that globally allele-bias cannot be highly deleterious to an individual. Instead, our results suggest that the allelic bias we observe is associated with common phenotypes and disease traits, suggesting that allelic bias in expression may contribute to phenotypic diversity and to risk for common diseases.

Methods

Sequence read alignment

The following description applies for the alignment of DNA Methylation, ChIP-Seq and DNase-Seq datasets. Single end sequencing data was mapped to a variant masked human reference genome (hg18) using Novoalign (www.novocraft.com). Unmapped and non-uniquely mapping reads were removed, and PCR duplicate reads were removed with Picard. Reads were processed with the Genome Analysis Toolkit (GATK)[30]. Specifically, reads underwent indel recalibration and variant realignment. Lastly, reads that overlapped with variant loci were split into the “p1” and “p2” allele according to whether the bases in the sequencing read matched the sequence from either the p1 or the p2 alleles.

For Hi-C datasets, read pairs were mapped independently to the variant masked genome using Novoalign. Reads were then manually paired using in house scripts. Non-uniquely mapping, unmapped reads, and PCR duplicate read pairs were removed. Reads pairs were then split into single reads and processed through the same GATK pipeline described above including indel re-alignment and variant recalibration. Finally, read pairs were manually re-paired using in house scripts. For mRNA-Seq, we mapped the paired-end data to a variant masked transcriptome using Novoalign.

Genotyping and haplotyping

Whole genome sequencing (WGS) data for the H1 genome were downloaded from the Sequence Read Archive Database (SRA049981). Reads were mapped to the hg18 reference using Novoalign. Unmapped and non-uniquely mapping reads were removed using in house scripts. PCR duplicate reads were removed using Picard[31]. The data was processed through the Genome Analysis Toolkit (GATK) best practices guidelines. We performed indel recalibration, variant realignment, variant calling using the Unified Genotyper, and variant recalibration was performed to achieve high quality genotyping.

Haplotyping was performed using the previously described HaploSeq method[22]. Briefly, Hi-C reads combined from each of the H1 derived lineages and whole genome sequencing were used as input sequencing into the HaploSeq algorithm in order to generate haplotype predictions. For final haplotype calls, Hi-C data was combined with WGS mate-pair data for the H1 genome. HapCUT generates several “blocks” for each chromosome. The vast majority of variants on each chromosome are in the “Most Variants Phased” (MVP) block. The MVP block for each chromosome was used as a “seed haplotype” for local conditional phasing using the Beagle v4.0[32]. This generates two haplotypes for each chromosome, one for the maternal allele and one for the paternal allele. Since we do not have information regarding the parent of origin in the H1 genome, we arbitrarily define each allele as the “p1” or “p2” allele (p1 and p2 for “parent 1” and “parent 2”). The p1 and p2 allele for different chromosomes are not necessarily derived from the same parent, as this information is only accessible if the sequence of H1’s parents were also available.

Identification of allelic genes

We considered the two replicates of mRNA-Seq data and used a negative binomial distribution (10% FDR) to calculate significantly biased genes between the two alleles, where genes are defined by merging isoforms (from RefSeq). Finally, we included only allelic genes that showed >35% MAF based on control sequencing datasets, such as DNA Methylation reads and genome sequencing.

Identification of allelic SNPs

We estimated if a SNP is allelic based on different types of readouts. In particular, we used ChIP-Seq, DHS, TF factor datasets independently to obtain readouts of each SNP between the two alleles. We then used a binomial statistic (with an expectation p=0.5) to identify significantly biased SNPs for a given dataset. FDR was based on 1000 random permutations. Lastly, we included only allelic SNPs that showed >35% MAF based on control sequencing datasets.

Identification of allelic methylation

We initially grouped CpGs around heterozygous variants and used a hyper-geometric test to evaluate significance and FDR was performed as described above.

Identification of allelic enhancers

To systematically study allelic enhancers, we combined several enhancer marks to obtain a combined acetylation bam file. This combined bam file gives us

the required coverage in an allelic context to perform an in-depth analyses. In particular, we combined data from H4K8ac, H4K91ac, H2BK120ac, H3K18ac, H3K23ac, H3K27ac, H3K4ac, H2AK5ac and H3K9ac marks. For evaluating allelic enhancers, we obtained readout for enhancers defined in Xie et al 20137 (±2.5kb from enhancer peaks) between the two alleles[4]. Then we used binomial to obtain significance at an FDR of 10%, as evaluated by the random permutation analyses (1000 permutations). By using acetylation alone, we identified 726 allelic enhancers. We performed similar analyses using DHS and identified 969 allelic enhancers, totaling to 1589 allelic enhancers. Similar to allelic SNPs and genes, we included allelic enhancers that showed >35% MAF based on control sequencing datasets.

Enhancer and gene annotations

The enhancer regions were defined as previously described[4]. Briefly, enhancer chromatin signatures were trained for p300 binding sites in H1 ES cells using RFECS algorithm based on H3K4me1, H3K4me3, and H3K27ac signals at 100bp bin size. Next, these modification signals in all cell lines were tested to predict enhancers. The predicted enhancers that overlap with H3K4me3 peaks or within 2.5kb of the transcription start site were removed. Enhancers were merged from all cell types if they are located close to each other (<2kb) by taking the midpoint at the center of the new enhancer. For the gene list, gene expression levels, house keeping genes, and lineage-specific genes we used the same data set as described in Xie et al[4]. For imprinting genes, we obtained 59 known

imprinting genes downloaded from publicly available imprinting gene database (http://www.geneimprint.com/).

Correlating allelic genes and allelic promoters

To investigate how many allelic gene promoter regions are consistent with allelic gene expression levels, first we selected allelic genes that contain at least one allelic SNP in their promoter regions (1.5kb upstream and downstream from transcription start site). We only considered allelic SNPs defined by DNaseI HS site, H3K4me3, histone Ac, combined H3K9me3 and H3K27me3, and DNA methylation because the functions of those chromatin marks at the promoter regions are well defined. If promoters are marked by allelic SNPs from H3K9me3/H3K27me3 or DNA methylation and the allelic gene expression levels are consistent with those promoter patterns, the genes can be explained by allelic repressive marks. If promoters are marked by allelic SNP from histone acetylations, H3K4me3, and DNaseI HS site and allelic gene expression levels are consistent with those promoter patterns, the genes can be explained by allelic active marks.

Identification of enhancer-promoter interactions

To investigate the linking between allelic genes and allelic enhancers we first defined enhancer-promoter interactions using Hi-C interaction frequency data. Hi-C interaction frequencies were calculated in terms of 5kb window and normalized using HiCNorm[33]. After that, we considered all pairs of promoters and

enhancers in each chromosome. Promoter regions were fixed as +/- 5kb surrounding transcriptional start sites and enhancer regions were defined by using different window sizes as 5kb, 10kb, 20kb, 30kb, 40kb, 50kb, 75kb, 100kb, 300kb, and 500kb surrounding center of each enhancer. The interaction frequencies between a promoter and an enhancer at a certain window size were calculated as (Interaction frequency / window size of an enhancer)*5kb. Final interaction scores were defined as summation of interaction frequencies between promoter and enhancer with multiple window sizes. To calculate significance of each enhancer-promoter interaction, we generated a random interaction frequency score by randomly permutated interaction frequencies between promoter and enhancer in each window size. The distribution of random interaction frequency scores was fit to Weibull distribution and p values of each interaction frequency between promoter and enhancer were calculated. At a p value cutoff of 1E-03, we defined enhancer-promoter interactions.

Correlation allelic gene and allelic enhancer

We calculated correlation coefficient between allelic gene and allelic enhancer. First we generate 1 by 10 vectors for allelic gene and allelic enhancer, respectively, for H1 and H1-derived four lineages. For each lineage, we assigned log2(p2 allele /p1 allele) and log2(p1 allele / p2 allele) values as allelic bias information. After constructing two 1 by 10 vectors for both allelic gene and allelic enhancer, we calculated the Pearson correlation coefficient between them.

4C-Seq analyses

Sequence reads were processes as follows. For each read, the first and second sequencing reads were checked to identify the presence of the primer sequences and any expected portion of the bait region. Any sequence with greater than 20% mismatches to the expected bait region was discarded. The reads were trimmed such that each read was represented as a 36-mer, with 20bp derived from the bait region and the subsequent 16bp, presumably containing the target region of interest.

4C-seq data was mapped to a version of the hg18 genome with known SNPs in the H1 genome masked to N, similar to other the strategy of mapping other sequence read datasets performed in this study. Custom indexes for this H1-masked hg18 genome were built using the 4cseqpipe “-build_re_db” command. The reads were mapped using the 4Cseqpipe software “-map” command to custom built indexes. Normalized contact intensities were derived using the 4seqpipe “-nearcis” command for a 1Mb region upstream and downstream of the bait locus. We then took the normalized fragment level interaction frequency tables and removed any fragments where a SNP either could create or disrupt a potential restriction enzyme site between the two alleles. In addition, given the short sequencing read length, any fragment with an insertion or deletion mapping within 16bp of the fragment end was removed. These final filtered sets of normalized fragment level interaction frequencies were then processed using a sliding window approach with the window size of 5kb and step size of 1kb using the average fragment interaction frequency over the 5kb

window. These sliding interaction frequency files were then quantile normalized across all replicates in order for comparison between experiments using the “normalize.quantiles.robust” function (with use.median=TRUE) in the “preprocessCore” library in R. For display purposes, the average of two replicates was converted to bedGraph format and displayed in the UCSC genome browser.

To identify regions that showed specific interactions with the bait region controlling for the genomic distance between loci, we developed a LOWESS regression model. We pooled the sliding window interaction frequency files from each of the 4C-seq replicates and performed LOWESS regression in R with the function “lowess” (with f=0.01) on the log-base10 transformed interaction frequencies controlling for the distance between the bait and potential interaction locus. We considered any region as showing “specific” interactions if it showed an increase in interaction frequency greater than 2.5 fold over expected given the distance between the bait and target loci. These were considered to be the “bait interacting regions.”

Figures and Tables

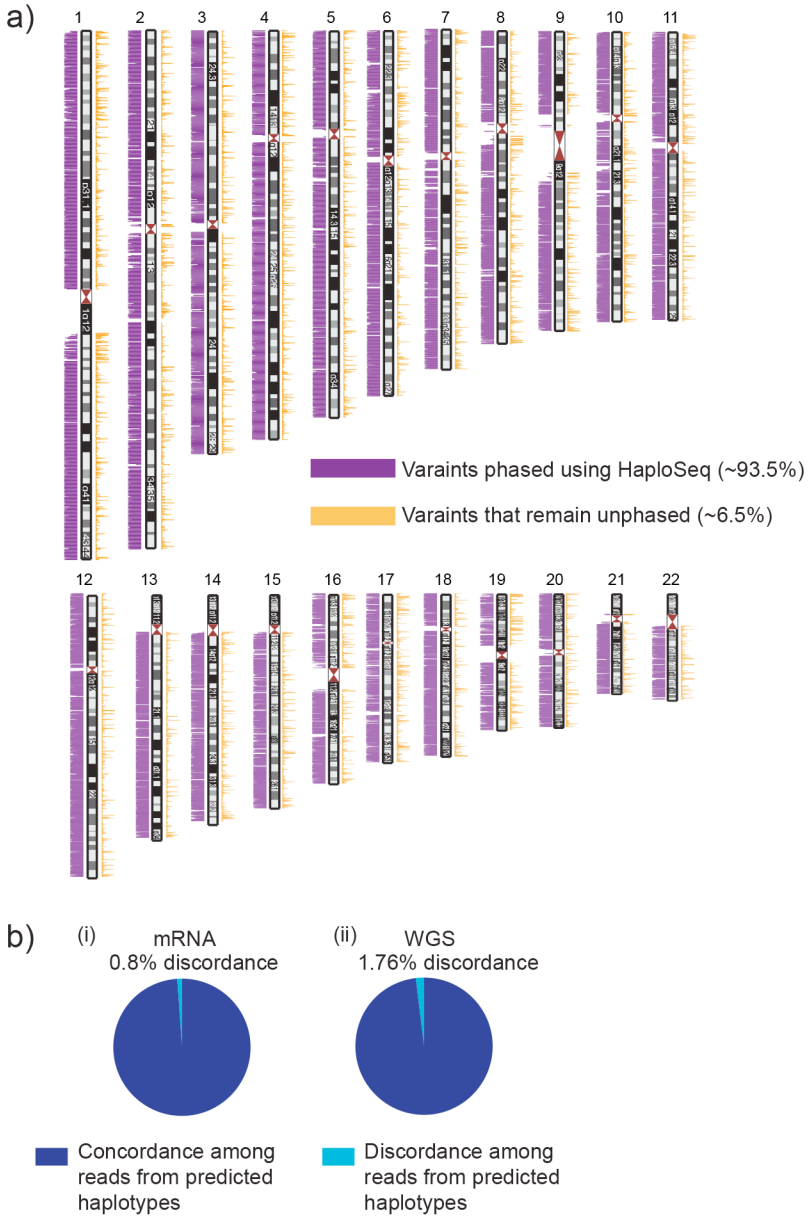

Figure 4-1: Haplotype phasing in H1.

a) Graph demonstrating HaploSeq phasing of variants for each chromosome. The left axis (purple) is the number of variants phased per 100kb bins and the right axis (gold) are the unphased variants. Over 93.5% alleles are phased using HaploSeq. b) Validation of haplotypes by (i) RNA-sequencing and (ii) whole-genome sequencing (WGS).

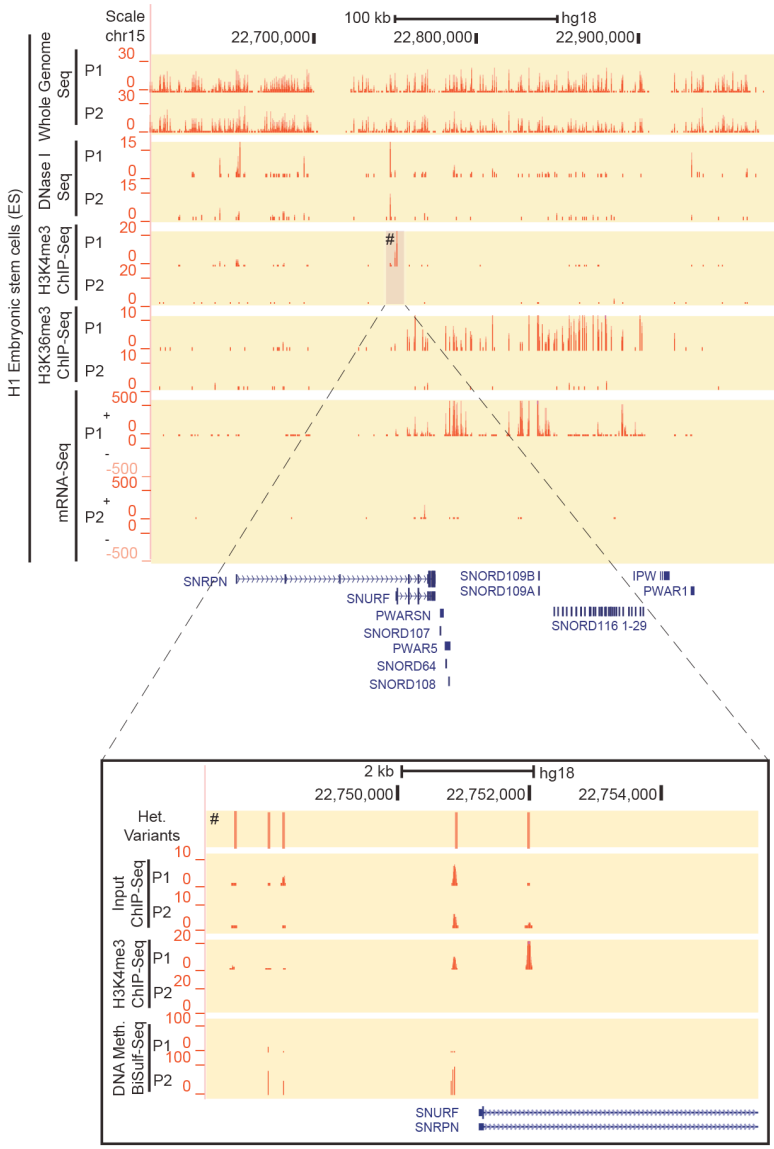

Figure 4-2: Recapitulating imprinting activity at SNRPN gene cluster.

Genome browser shots of allele specific DNaseI Hypersensitivity, chromatin modifications, and mRNA-sequencing. The two parental alleles are designated as P1 and P2. For mRNA-seq, data is shown in a strand-specific manner as well. Inset labeled with # from panel c showing mutually exclusive allele specific DNA-methylation and H3K4me3 at the SNURF promoter at the SNRPN gene cluster. Together, we observe functional activity only in the P1 allele while P2 allele is non-functional due to methylated and therefore inactive P2 promoter.

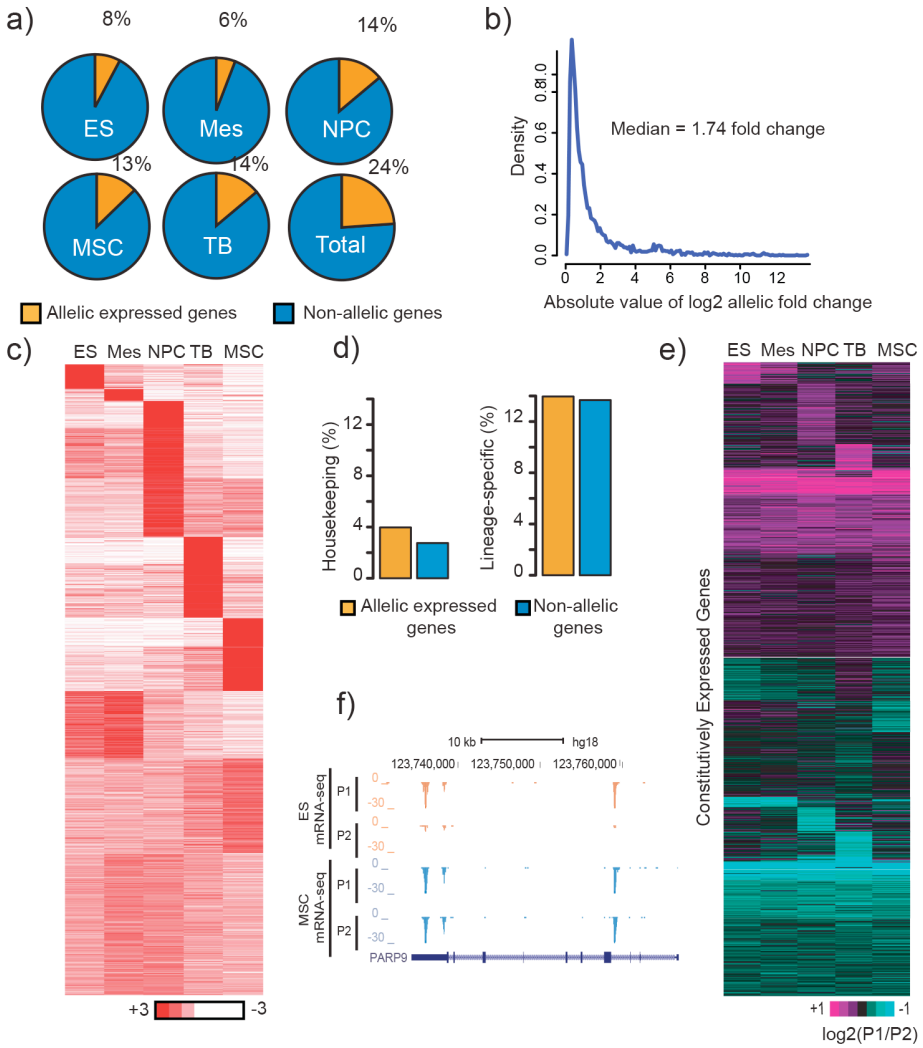

Figure 4-3: Widespread allele specific gene-expression.

a) Pie charts showing the proportion of genes with detectable allelic expression that show statistically significant allelic bias in each lineage. b) Density plot of the absolute value of the fold change in expression between alleles (log 2). c) Heat map showing K-means clustering (k=12) of gene expression levels of allele biased genes across each of the 5 H1 hESC derived lineages. The expression levels are shown as the fold-change of expression in each lineage relative to the average expression level across each of the 5 lineages. d) Fraction of housekeeping genes, and lineage-restricted genes that show allele biased expression. e) Heat map showing k-means (k=20) clustering of allelic expression ratios at the genes with constitutive expression in each of the 5 lineages. f) UCSC genome browser of PARP9 showing allelic bias favoring the p1 allele in ES cells while it shows no allelic bias in MSC despite similar expression levels.

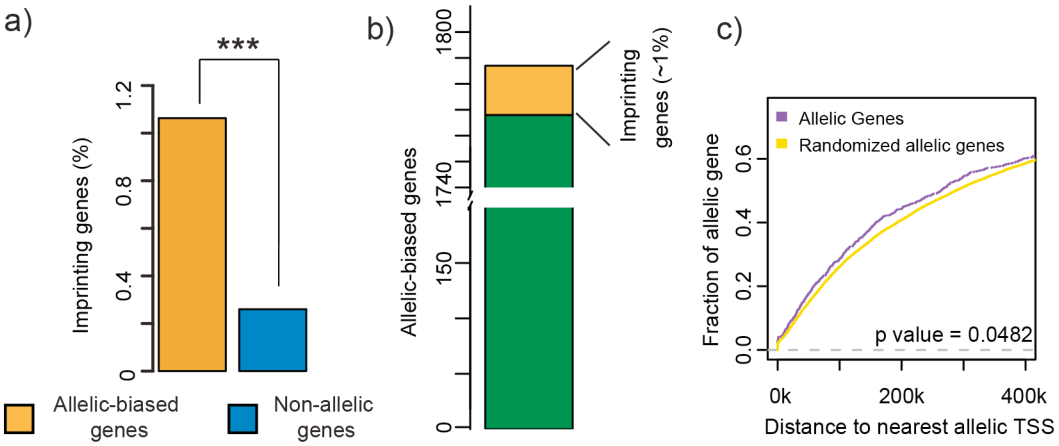

Figure 4-4: Allelic bias is enriched among imprinted genes.

a) Allelic biases are enriched in imprinted genes (p value 1.3E-5). b) Fraction of imprinted genes among allelic genes. c) Empirical cumulative density plot of the distance between each allele-biased gene and the nearest allele-biased gene (purple) as compared with randomly chosen genes (yellow). The difference from an allele-biased gene to the nearest allele-biased gene is less than what would be expected at random (p=0.0482, Wilcoxan rank sum test), however, the difference is subtle, indicating that most allele biased expression does not occur in clusters.

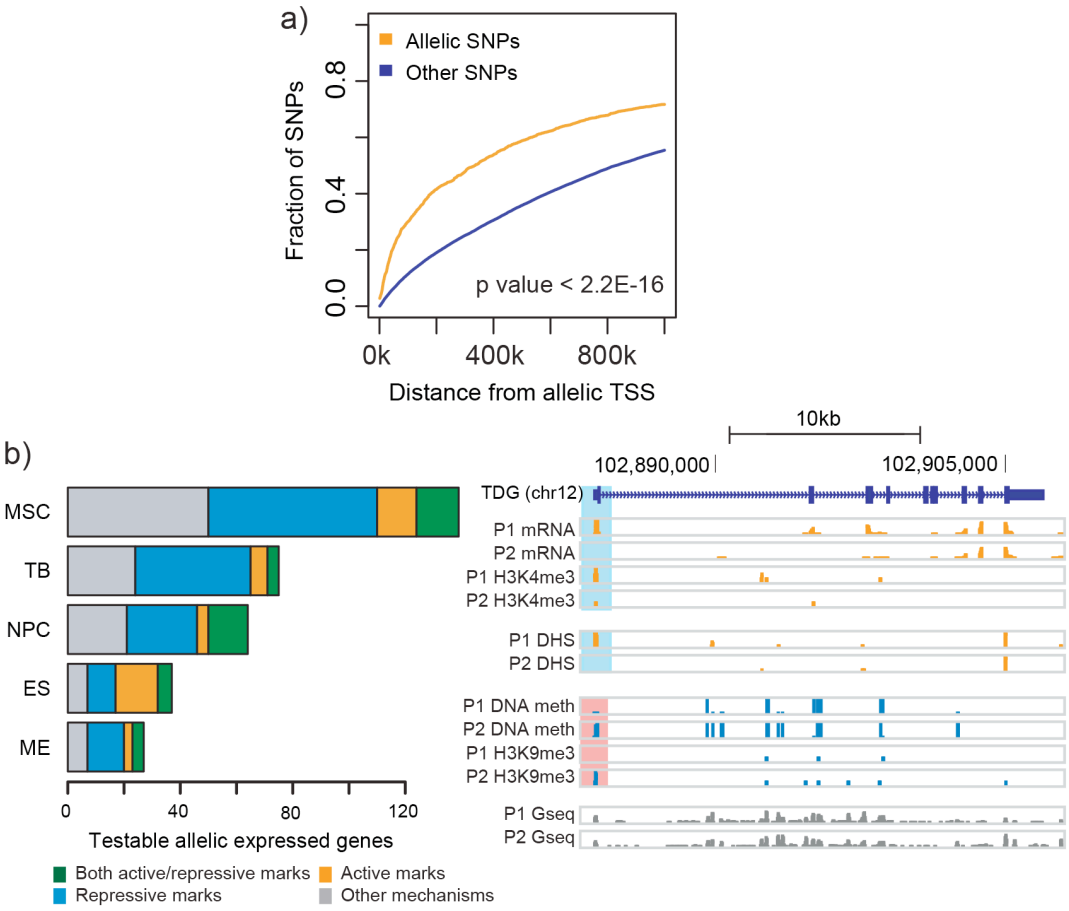

Figure 4-5: Allelic promoter bias correlates with allelic transcription.

a) Empirical cumulative density plot of distances from allelic SNPs and non-allelic SNPs to the nearest allele-specific gene transcription start site. Allele specific SNPs are defined using histone acetylation, combined H3K9me3/H3K27me3, DNaseI HS, and H3K4me3. Allele specific SNPs tend to be located closer to allele specific genes compared with non-allele specific SNPs (<2.2E-16 KS-test). b) Number of allele specific genes showing consistent allele specific chromatin states in their promoter regions. Allele specific SNPs identified by H3K4me3, DNaseI HS, and histone acetylation are considered as active allele specific SNPs and those identified by DNA methylation and H3K9me3/27me3 are considered as inactive allele specific SNPs. c) Example UCSC genome browser shot of the TDG gene, an allele biased gene. Allele specific chromatin states at promoter regions are consistent with allele specific gene expression levels.

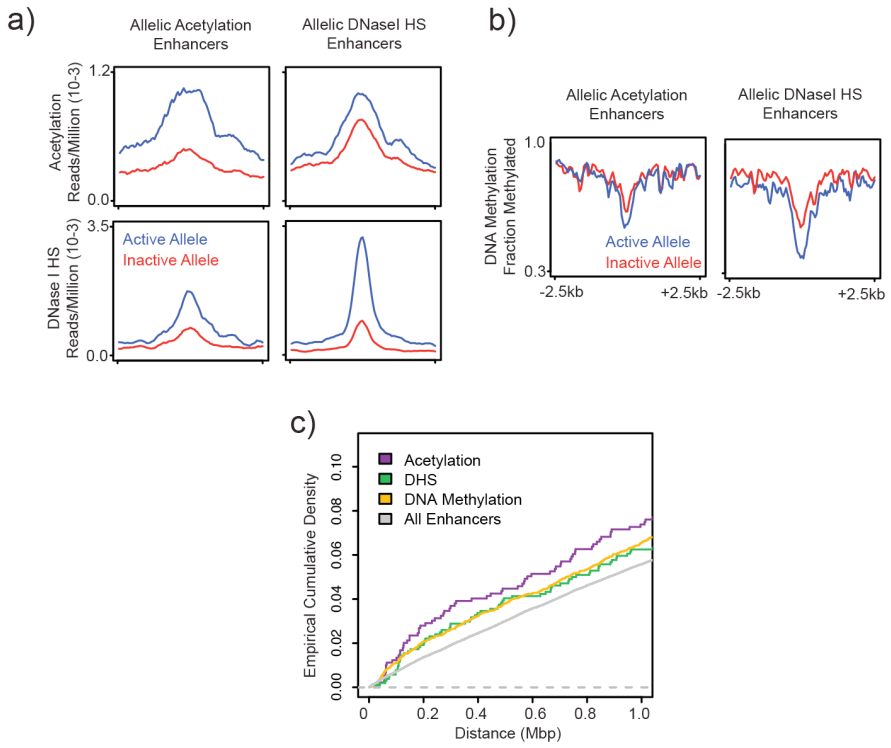

Figure 4-6: Patterns of allelic enhancer sites.

a) Plots demonstrating the enrichment of acetylation (top row), DNase I HS (bottom row). b) Reduced DNA Methylation activity at allelic enhancers as defined by histone acetylations and DNase I HS. c) Enhancers that display allelic activity bias tend to be closer to allele specific genes. The distance between allelic genes and enhancers as defined by allelic acetylation (purple), DNaseI HS (green), DNA methylation (yellow) or all enhancers (blue) are shown.

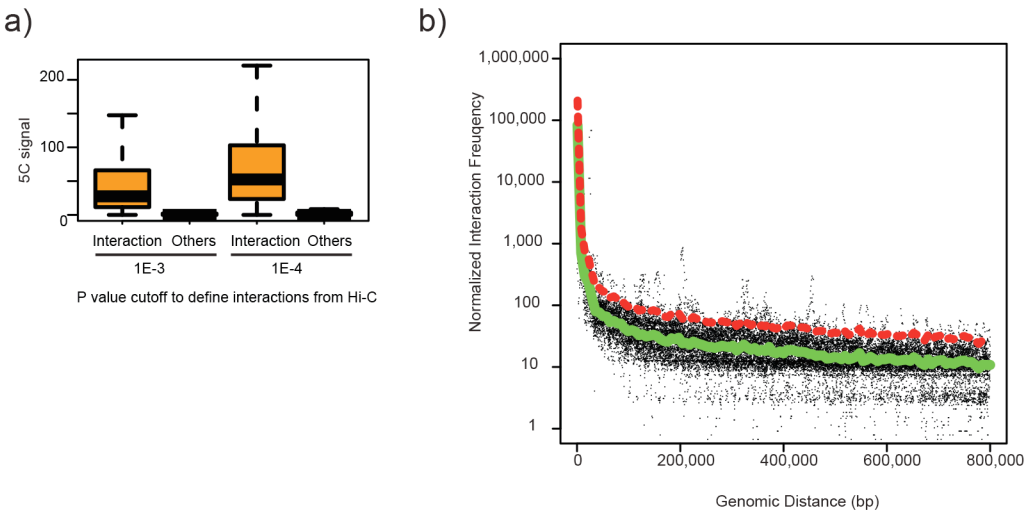

Figure 4-7: Spatial proximity estimates based on Hi-C and 4C-Seq are of high quality.

a) Distribution of 5C signals between interacting pairs ('Interaction') and non-interacting pairs ('Others') defined by Hi-C interaction frequency score from our method at different pvalue cutoffs. Regardless of pvalue cutoff, we observe a strong correlation among these two predictions and therefore validating our Hi-C based interaction scores. b) Scatter plot of LOWESS regression of 4C-seq data. The x-axis shows the genomic distance between the bait region and the putative target region. The y-axis is the log base-10 of the quantile normalized interaction frequencies. LOWESS was performed to generate an expected interaction frequency at each genomic distance (green line). A cut off of 2.5 fold over expected (shown in the red dashed line) is used to determine if a region shows specific interactions.

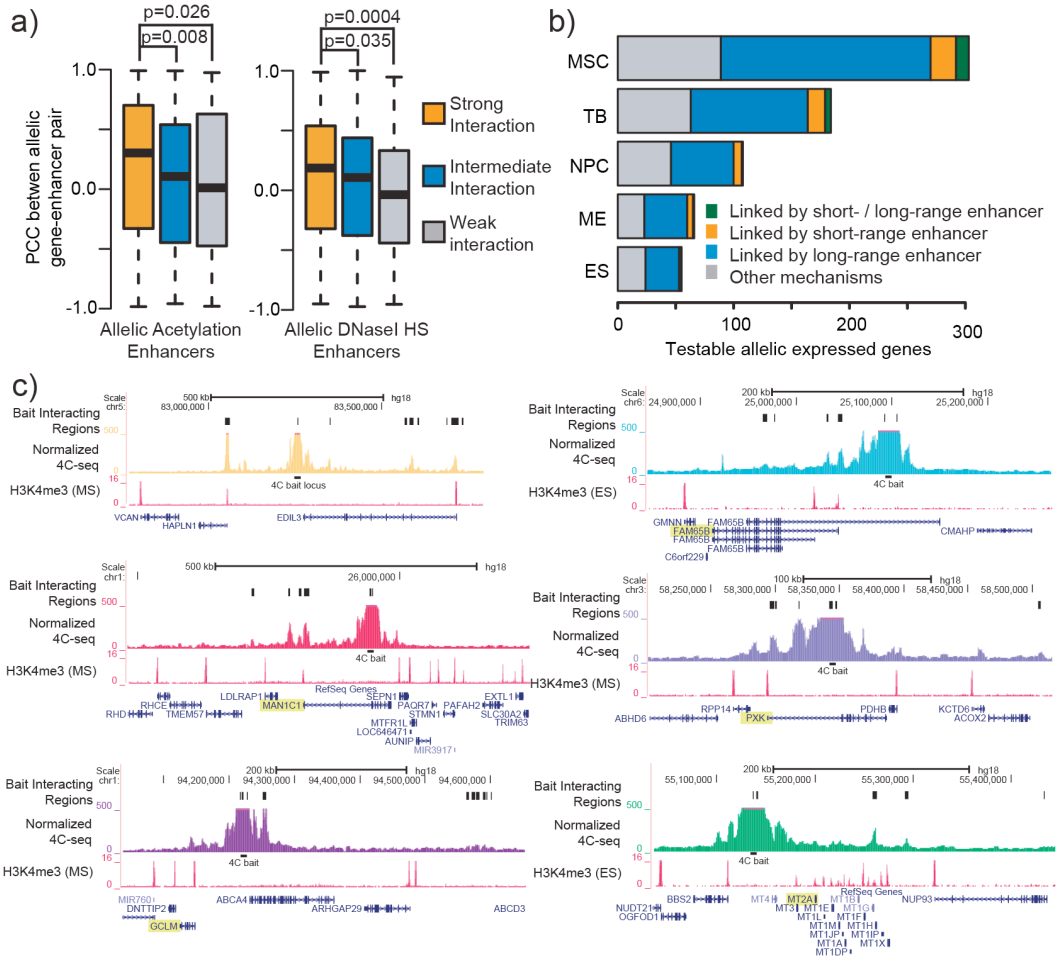

Figure 4-8: Spatially proximal allelic enhancers correlate to transcription.

a) The Pearson correlation coefficients of allelic biased gene-enhancer pair activities across five cell types. Allelic biased gene-enhancer pairs are grouped into strongly (top 30%), weakly (bottom 30%), and intermediately interacting pairs. b) Number of allele specific genes linked by allele specific enhancers. The long-range enhancer-promoter interactions are defined using Hi-C interaction frequencies. The short-range enhancer-promoter interactions are any enhancers <20kb from TSS. If allele specific gene expression patterns are consistent with allele specific enhancer activities interactions, they are shown here. c) Normalized 4C-seq interaction frequencies surrounding a bait region located in the 6 allelic enhancers. Regions with significant interactions according to the LOWESS model are marked “Bait interacting Regions.” While GCLM interaction falls less than the LOWESS threshold, no specific interactions between the enhancer and the MT2A gene is found.

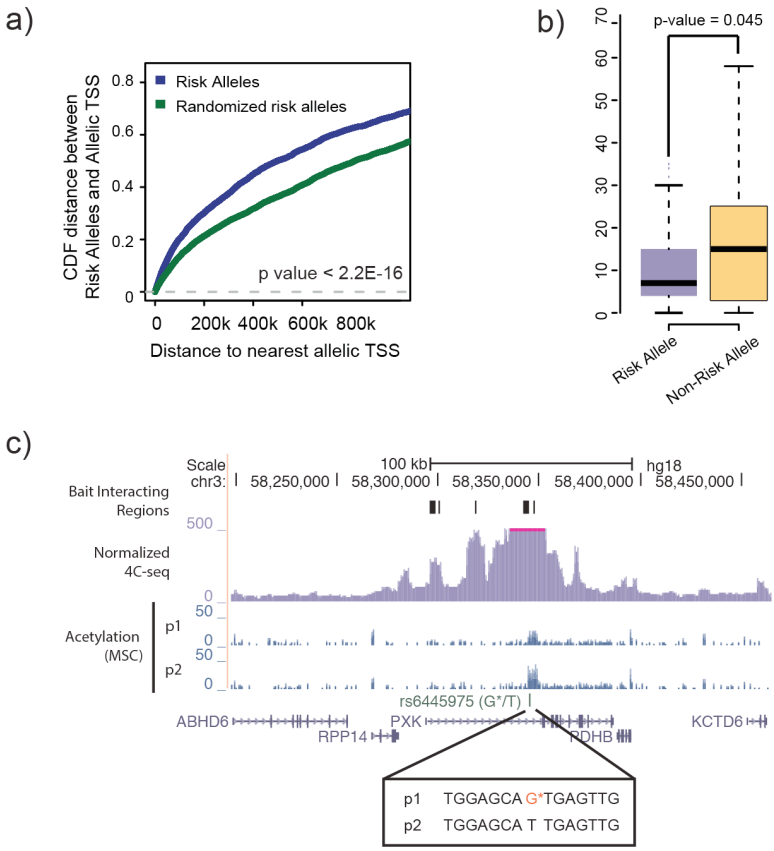

Figure 4-9: Allelic bias may contribute to human disease.

a) Empirical cumulative density plots of the distance between SNPs and allelic genes. SNPs are either categorized as GWAS associated if the SNP is in the GWS catalog or is in LD (r2 > 0.8) with a SNP in the GWAS catalog. As a control, an equal number of randomly selected SNPs (and SNPs in LD with the random selection) were used for comparison. b) Chromatin activity over risk and non-risk alleles in H1. For each SNP in the above-mentioned GWAS catalog, we calculated the number of reads from active chromatin marks (histone acetylation, DHS, H3K4me3, H3K4me1, H3K36me3) on the risk and non-risk alleles. c) Normalized 4C-seq interaction frequencies from an allele biased enhancer located in the PXK gene. The enhancer shows specific interactions with the promoter of the PXK gene. In addition, the H1 genome has a SNP located in this allele biased enhancer that has been previously linked to Systemic Lupus Erythematosis. At this enhancer, the risk allele (labeled with an asterisk) is associated with reduced enhancer activity as measured by acetylation levels compared with the non-risk allele.

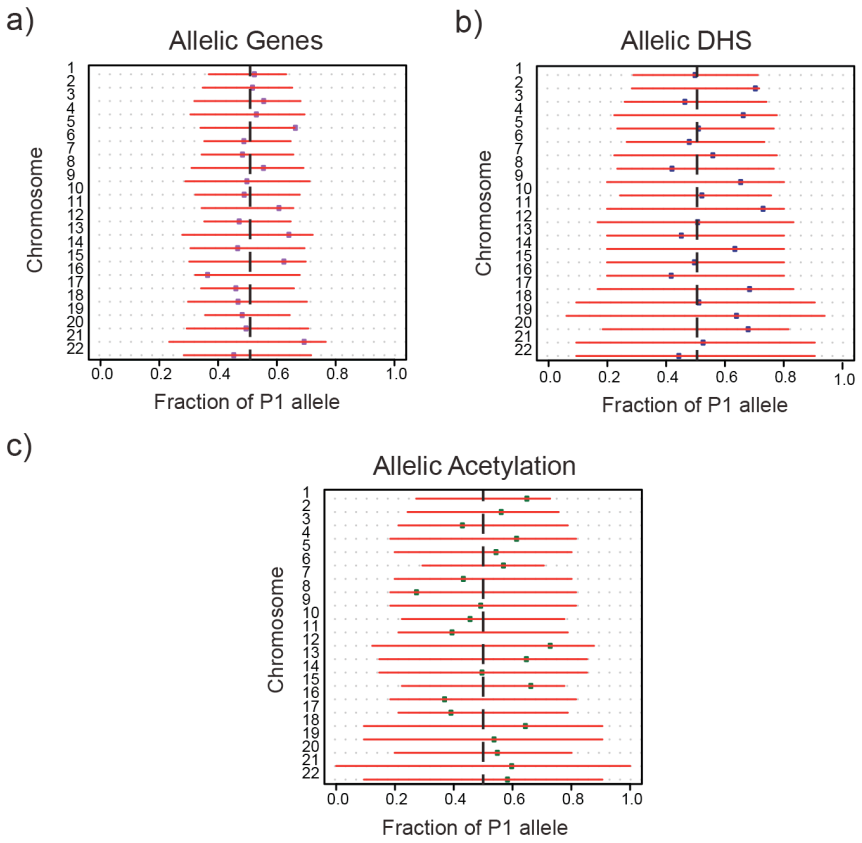

Figure 4-10: Allelic activity occur from both parental haplotypes.

Fraction of P1 allelic genes (a), DHS (b) and acetylaiton (c) per chromosome averaged accross 5 cell-types (dots). As the number of allelic events are lesser per chromosome, they might appear as a bias towards one parent. However, 95% bayes binomial confidence interval based on the number of allelic events, shows that there is no significant deviaiton of allelic activities towards any one parent and that the observed deviation could be explained by expected deviaiton from low counts of allelic events.

Table 4-1: Number of reads in the Hi-C experiment.

Table depicting read counts in Hi-C datasets accross 5 lineages. Together, we have ~3.85 billion reads.

Cell Type	Replicate	Total Reads	Short reads (<500bp)	Percent	cis reads	Percent	Trans reads	Percent
ES	rep1	331,587,795	122,070,404	36.81%	159,920,890	48.23%	49,596,501	14.96%
	rep2	743,132,905	276,597,229	37.22%	233,301,469	31.39%	233,234,207	31.39%
ME	rep1	527,651,650	218,062,039	41.33%	209,163,061	39.64%	100,426,550	19.03%
	rep2	329,765,028	86,437,620	26.21%	205,629,844	62.36%	37,697,564	11.43%
MSC	rep1	273,900,059	46,155,555	16.85%	188,679,257	68.89%	39,065,247	14.26%
	rep2	324,325,610	53,579,351	16.52%	221,655,061	68.34%	49,091,198	15.14%
NPC	rep1	259,265,402	73,500,282	28.35%	56,278,964	21.71%	129,486,156	49.94%
	rep2	361,610,256	101,875,277	28.17%	75,293,949	20.82%	184,441,030	51.01%
TB	rep1	409,236,714	218,716,568	53.45%	100,776,836	24.63%	89,743,310	21.93%
	rep2	297,555,667	63,254,013	21.26%	117,315,748	39.43%	116,985,906	39.32%

Acknowledgements

Chapter 4, in full, has been submitted for review in Nature. Jesse R Dixon*, Inkyung Jung*, Siddarth Selvaraj*, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Victor V Lobanenkov, Joseph Ecker, James Thomson, Bing Ren. “Global Reorganization of Chromatin Architecture during Embryonic Stem Cell Differentiation”. The dissertation author was a primary investigator and author of this paper. In particular, the dissertation author performed haplotyping of H1 cells and analyzed gene regulation patterns in an allele-specific manner.

References

1. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G.H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Jonasdottir, A., Styrkarsdottir, U., Gretarsdottir, S., Magnusson, K.P., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H.G., Stefansson, T., Leifsson, B.G., Thorsteinsdottir, U., Lamb, J.R., Gulcher, J.R., Reitman, M.L., Kong, A., Schadt, E.E. & Stefansson, K. Genetics of gene expression and its effect on disease. Nature 452, 423-428 (2008).

2. Ben-Porath, I., Thomson, M.W., Carey, V.J., Ge, R., Bell, G.W., Regev, A. & Weinberg, R.A. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nature genetics 40, 499-507 (2008).

3. Schnabel, M., Marlovits, S., Eckhoff, G., Fichtel, I., Gotzen, L., Vecsei, V. & Schlegel, J. Dedifferentiation-associated changes in morphology and gene expression in primary human articular chondrocytes in cell culture. Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society 10, 62-70 (2002).

4. Xie, W., Schultz, M.D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J.W., Tian, S., Hawkins, R.D., Leung, D., Yang, H., Wang, T., Lee, A.Y., Swanson, S.A., Zhang, J., Zhu, Y., Kim, A., Nery, J.R., Urich, M.A., Kuan, S.,

Yen, C.A., Klugman, S., Yu, P., Suknuntha, K., Propson, N.E., Chen, H., Edsall, L.E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W.Y., Chi, N.C., Antosiewicz-Bourget, J.E., Slukvin, I., Stewart, R., Zhang, M.Q., Wang, W., Thomson, J.A., Ecker, J.R. & Ren, B. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell 153, 1134-1148 (2013).

5. Zhu, J., Adli, M., Zou, J.Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P.L., Bennett, D.A., Houmard, J.A., Muoio, D.M., Onder, T.T., Camahort, R., Cowan, C.A., Meissner, A., Epstein, C.B., Shoresh, N. & Bernstein, B.E. Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell 152, 642-654 (2013).

6. Gifford, C.A., Ziller, M.J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A.K., Kelley, D.R., Shishkin, A.A., Issner, R., Zhang, X., Coyne, M., Fostel, J.L., Holmes, L., Meldrim, J., Guttman, M., Epstein, C., Park, H., Kohlbacher, O., Rinn, J., Gnirke, A., Lander, E.S., Bernstein, B.E. & Meissner, A. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell 153, 1149-1163 (2013).

7. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B. & Ecker, J.R. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315-322 (2009).

8. Maunakea, A.K., Chepelev, I. & Zhao, K. Epigenome mapping in normal and disease States. Circulation research 107, 327-339 (2010).

9. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. & Bernstein, B.E. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43-49 (2011).

10. Bird, A. DNA methylation patterns and epigenetic memory. Genes & development 16, 6-21 (2002).

11. Jackson, M., Krassowska, A., Gilbert, N., Chevassut, T., Forrester, L., Ansell, J. & Ramsahoye, B. Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. Molecular and cellular biology 24, 8862-8871 (2004).

12. Tsumura, A., Hayakawa, T., Kumaki, Y., Takebayashi, S., Sakaue, M., Matsuoka, C., Shimotohno, K., Ishikawa, F., Li, E., Ueda, H.R., Nakayama, J. & Okano, M. Maintenance of self-renewal ability of mouse embryonic stem cells in

the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Genes to cells : devoted to molecular & cellular mechanisms 11, 805-814 (2006).

13. Kouzarides, T. Chromatin modifications and their function. Cell 128, 693-705 (2007).

14. Yao, T.P., Oh, S.P., Fuchs, M., Zhou, N.D., Ch'ng, L.E., Newsome, D., Bronson, R.T., Li, E., Livingston, D.M. & Eckner, R. Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300. Cell 93, 361-372 (1998).

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

16. Kosak, S.T. & Groudine, M. Form follows function: The genomic organization of cellular differentiation. Genes & development 18, 1371-1384 (2004).

17. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. Nature 489, 109-113 (2012).

18. Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., Sim, H.S., Peh, S.Q., Mulawadi, F.H., Ong, C.T., Orlov, Y.L., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C.L., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K.I., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M.J., Cheung, E., Liu, E., Sung, W.K., Snyder, M. & Ruan, Y. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84-98 (2012).

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

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

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

22. Selvaraj, S., J, R.D., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nature biotechnology 31, 1111-1118 (2013).

23. Zeschnigk, M., Schmitz, B., Dittrich, B., Buiting, K., Horsthemke, B. & Doerfler, W. Imprinted segments in the human genome: different DNA methylation patterns in the Prader-Willi/Angelman syndrome region as determined by the genomic sequencing method. Human molecular genetics 6, 387-395 (1997).

24. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R. & Chess, A. Widespread monoallelic expression on human autosomes. Science 318, 1136-1140 (2007).

25. Heinz, S., Romanoski, C.E., Benner, C., Allison, K.A., Kaikkonen, M.U., Orozco, L.D. & Glass, C.K. Effect of natural genetic variation on enhancer selection and function. Nature 503, 487-492 (2013).

26. Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., Li, J., Xie, D., Olarerin-George, A., Steinmetz, L.M., Hogenesch, J.B., Kellis, M., Batzoglou, S. & Snyder, M. Extensive variation in chromatin states across humans. Science 342, 750-752 (2013).

27. Kilpinen, H., Waszak, S.M., Gschwind, A.R., Raghav, S.K., Witwicki, R.M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N.I., Yurovsky, A., Lappalainen, T., Romano-Palumbo, L., Planchon, A., Bielser, D., Bryois, J., Padioleau, I., Udin, G., Thurnheer, S., Hacker, D., Core, L.J., Lis, J.T., Hernandez, N., Reymond, A., Deplancke, B. & Dermitzakis, E.T. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science 342, 744-747 (2013).

28. van de Werken, H.J., Landan, G., Holwerda, S.J., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A., Verstegen, M.J., de Wit, E., Tanay, A. & de Laat, W. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nature methods 9, 969-972 (2012).

29. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., & Parkinson, H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research 42, D1001-D1006 (2014).

30. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D. & Daly, M.J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 43, 491-498 (2011).

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

32. Browning, B.L. & Browning, S.R. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. Genetics 194, 459-471 (2013).

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

Chapter 5: Perspectives on utility of 3D genome information

The eukaryotic genome has a non-random spatial organization, facilitating and coordinating diverse cellular processes such as DNA Replication and transcription[1-5]. Methods based on Fluorescent in-situ hybridization (FISH)[6], X-Ray tomography[7], and chromosome conformation capture (3C)[8] have revealed multiple aspects of genome structure – chromosome territories (CT)[4,5,7], compartments of active and inactive chromatin[1,4], and physical interactions governed at individual loci revealing long range chromatin looping between genes and regulating elements[3,9]. Nevertheless, these methods are low-throughput and therefore not amenable to genome-wide analyses of the 3D genome. With the invention of Hi-C[10], large-scale, systematic studies of genome structure are now possible. Characterizing the genome sequence, and structure, along with gene expression and the epigenome, will further our understanding on how cell-type specific gene regulation is achieved and in elucidating its dynamic nature through cellular differentiation. In this chapter, I will discuss current and future prospective utilities of obtaining 3D genome information.

First-generation maps of 3D genome have suggested that the genome is organized in to topologically associated domains (TADs)[11-13]. In this thesis, I have described a computational strategy to identify TADs using Hi-C datasets[11]. I have also shown that TADs are pervasive across the genome and are highly conserved between human and mouse, suggesting an evolutionary aspect to TADs and genome structure. While TAD locations are stable across cell-types, sub-TAD chromatin interactions can alter their 3D shape, driving cell-type specific gene regulation. Indeed, we observed that chromatin interactions

enriched in mouse embryonic cells (ES), in comparison to mouse cortex, were enriched for ES specific genes. However, our analyses were constrained by low resolution Hi-C datasets. Recently, Phillips-Cremins and colleagues[14] have generated high-resolution chromatin maps to reveal dynamic changes in TAD shapes correlating with gene expression. To this end, ~90% of disease-associated sequence variants reside in non-coding regulatory sequences with unknown target genes[15,16]. Consequently, mapping chromatin interactions between variants and promoters can help identify disease-associated target genes. While, we and others have established the role of genome structure on its function[1,9,11,12,14], a predictive model for gene regulation that underlies contributions of genome sequence, structure, and epigenome is yet to be performed. With recent developments in genome editing tools such as TALEN[17,18], and CRISPR[19,20], it is possible to perturb regulatory sequences or TAD boundaries, offering a way to investigate the contribution of genome sequence on its structure and function.

While genome-editing methods can perturb genetic sequences in an elegant manner, they are still low-throughput. By exploiting the sequence differences between the two homologous chromosomes in the diploid human or mouse genomes, we can potentially correlate these to changes in structure and regulation. Nevertheless, such an analysis requires the knowledge of long-range haplotypes or “phasing”, which has long remained an elusive goal[21,22]. In this thesis, I invented HaploSeq, which builds on the Hi-C protocol and offers a rigorous solution for generating chromosome-scale phasing[23]. I demonstrated

HaploSeq in two systems, CASTxJ129 mouse cells and human GM12878 cells, for which genotyping of parent-child trio generated haplotype information a prioiri. While HaploSeq phased ~95% of alleles in mouse, I coupled HaploSeq with local-conditional phasing to obtain high-resolution haplotypes in low variant density human cells. Several future directions can strengthen haplotyping capabilities of Haploseq. For one, concurrent genotyping and haplotyping from Hi-C datasets can generate complete genetic makeup of an individual from a single assay. Second, phasing structural variants can help in understanding disease states such as cancer progression[24] and autism[25,26], where large insertions, inversions, and deletions are known to play a disruptive role. Third, recent developments might extend HaploSeq to phase polyploid agricultural crops[27]. In addition, Job Dekker, Jay Shendure and colleagues[28,29] have demonstrated de novo assembly capabilities of Hi-C datasets. Taken together, Hi-C is emerging to be a multi-purpose tool, revealing several unique aspects of genome sequence, and structure.

Recently, the Roadmap epigenome consortium has generated comprehensive profiles of DNA methylation, histone modifications, chromatin accessibility, and gene expression across H1 human embryonic stem cells (ES) and four ES-derived lineages, to explore gene regulation patterns across differentiation[30,31]. As our lab performed Hi-C on each of these 5 lineages, I performed HaploSeq to phase 93.5% of the variants to obtain chromosome-scale haplotypes. Utilizing the haplotype-resolved genome, we analyzed changes in genome structure and epigenome to correlate with gene regulation patterns

through the differentiation process. Our analyses revealed ~24% of allelic genes and such allelic transcription correlated with allelic chromatin states of promoters and enhancers, as well as supported by chromatin interactions from Hi-C. While we performed the first integrative analyses of genome sequence, structure and epigenome to decipher gene regulation patterns, the sparse number of variants (SNPs) in humans did not allow comprehensive predictive modeling of these extensive datasets. As a future prospect, we plan to recapitulate the above-mentioned analyses in the haplotype-resolved CASTxJ129 mouse system, as it contains 7-10× more variants than humans. Specifically, such a system can potentially allow us to detect allelic activity of many more functional elements, enabling detailed analyses of allelic gene regulation.

Altogether, sequencing methods that profile 3D genome information in an unbiased fashion such as Hi-C, not only can inform target genes for non-coding regulatory sequences, but also which of the two alleles are interacting. At present, Hi-C uses a single restriction enzyme to digest chromatin and consequently generates 3D genome data that is biased towards the location of restriction enzyme cut sites used. In principle, Hi-C can be performed with multiple restriction enzymes and this can potentially achieve a more uniform coverage of the genome, enabling a more complete analysis of 3D genome, haplotypes, as well as de novo assembly. Moreover, recent advancements on the lines of single-cell Hi-C[32], and targeted chromatin conformation based capture-C[33] will provide novel aspects of 3D genome at a higher resolution and undoubtedly will take us to closer to understanding of how cells read and

interpret genetic information. Furthermore, the recent NIH-RFI on 3D-nucleome might enable extensive profiling of genome structure and epigenome datasets among distinct individuals and conditions, allowing for better understanding of gene regulation in development and disease.

References

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

2. Kosak, S.T. & Groudine, M. Form follows function: The genomic organization of cellular differentiation. Genes & development 18, 1371-1384 (2004).

3. Chambeyron, S. & Bickmore, W.A. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. Genes & development 18, 1119-1130 (2004).

4. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature reviews. Genetics 2, 292-301 (2001).

5. Cremer, T., Cremer, M., Dietzel, S., Muller, S., Solovei, I. & Fakan, S. Chromosome territories--a functional nuclear landscape. Current opinion in cell biology 18, 307-316 (2006).

6. Solovei, I., Cavallo, A., Schermelleh, L., Jaunin, F., Scasselati, C., Cmarko, D., Cremer, C., Fakan, S. & Cremer, T. Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). Experimental cell research 276, 10-23 (2002).

7. Clowney, E.J., LeGros, M.A., Mosley, C.P., Clowney, F.G., Markenskoff-Papadimitriou, E.C., Myllys, M., Barnea, G., Larabell, C.A. & Lomvardas, S. Nuclear aggregation of olfactory receptor genes governs their monogenic expression. Cell 151, 724-737 (2012).

8. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. Science 295, 1306-1311 (2002).

9. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. Nature 489, 109-113 (2012).

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

11. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376-380 (2012).
12. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Bluthgen, N., Dekker, J. & Heard, E. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381-385 (2012).

13. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. & Cavalli, G. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 148, 458-472 (2012).

14. Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y., Bland, M.J., Wagstaff, W., Dalton, S., McDevitt, T.C., Sen, R., Dekker, J., Taylor, J. & Corces, V.G. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell 153, 1281-1295 (2013).

15. Heinz, S., Romanoski, C.E., Benner, C., Allison, K.A., Kaikkonen, M.U., Orozco, L.D. & Glass, C.K. Effect of natural genetic variation on enhancer selection and function. Nature 503, 487-492 (2013).

16. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. & Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106, 9362-9367 (2009).

17. Cade, L., Reyon, D., Hwang, W.Y., Tsai, S.Q., Patel, S., Khayter, C., Joung, J.K., Sander, J.D., Peterson, R.T. & Yeh, J.R. Highly efficient generation of heritable zebrafish gene mutations using homo- and heterodimeric TALENs. Nucleic acids research 40, 8001-8010 (2012).

18. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. & Charpentier, E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337, 816-821 (2012).

19. Jinek, M., East, A., Cheng, A., Lin, S., Ma, E. & Doudna, J. RNA-programmed genome editing in human cells. eLife 2, e00471 (2013).

20. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. & Church, G.M. RNA-guided human genome engineering via Cas9. Science 339, 823-826 (2013).

21. Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J. & Schork, N.J. The importance of phase information for human genomics. Nature reviews. Genetics 12, 215-223 (2011).
22. Browning, S.R. & Browning, B.L. Haplotype phasing: existing methods and new developments. Nature reviews. Genetics 12, 703-714 (2011).

23. Selvaraj, S., J, R.D., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nature biotechnology 31, 1111-1118 (2013).

24. Yang, L., Luquette, L.J., Gehlenborg, N., Xi, R., Haseley, P.S., Hsieh, C.H., Zhang, C., Ren, X., Protopopov, A., Chin, L., Kucherlapati, R., Lee, C. & Park, P.J. Diverse mechanisms of somatic structural variations in human cancer genomes. Cell 153, 919-929 (2013).

25. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell 148, 1223-1241 (2012).

26. Michaelson, J.J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., Wu, W., Corominas, R., Peoples, A., Koren, A., Gore, A., Kang, S., Lin, G.N., Estabillo, J., Gadomski, T., Singh, B., Zhang, K., Akshoomoff, N., Corsello, C., McCarroll, S., Iakoucheva, L.M., Li, Y., Wang, J. & Sebat, J. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 151, 1431-1442 (2012).

27. Berger, E., Yorukoglu, D., Peng, J. & Berger, B. HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data. PLoS computational biology 10, e1003502 (2014).

28. Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. & Shendure, J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature biotechnology 31, 1119-1125 (2013).

29. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from in vivo DNA interaction frequency. Nature biotechnology 31, 1143-1147 (2013).

30. Xie, W., Schultz, M.D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J.W., Tian, S., Hawkins, R.D., Leung, D., Yang, H., Wang, T., Lee, A.Y.,

Swanson, S.A., Zhang, J., Zhu, Y., Kim, A., Nery, J.R., Urich, M.A., Kuan, S., Yen, C.A., Klugman, S., Yu, P., Suknuntha, K., Propson, N.E., Chen, H., Edsall, L.E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W.Y., Chi, N.C., Antosiewicz-Bourget, J.E., Slukvin, I., Stewart, R., Zhang, M.Q., Wang, W., Thomson, J.A., Ecker, J.R. & Ren, B. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell 153, 1134-1148 (2013).

31. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S. & Thomson, J.A. The NIH Roadmap Epigenomics Mapping Consortium. Nature biotechnology 28, 1045-1048 (2010).

32. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A. & Fraser, P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502, 59-64 (2013).

33. Hughes, J.R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R. & Higgs, D.R. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nature genetics 46, 205-212 (2014).