Functional Organization of the Maternal and Paternal Human 4D Nucleome

Stephen Lindsly[1], Wenlong Jia[2], Haiming Chen[1], Sijia Liu[3], Scott Ronquist[1], Can Chen[4,7], Xingzhao Wen[5], Gabrielle A. Dotson[1], Charles Ryan[1,8,9] Gilbert S. Omenn[1], Shuai Cheng Li[2], Max Wicha[1,6], Alnawaz Rehemtulla[6], Lindsey Muir[1], Indika Rajapakse[1,7,*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor

[2]Department of Computer Science, City University of Hong Kong

[3]MIT-IBM Watson AI Lab, IBM Research

[4]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor

[5]Department of Bioengineering, University of California San Diego, La Jolla

[6]Department of Hematology/Oncology, University of Michigan, Ann Arbor

[7]Department of Mathematics, University of Michigan, Ann Arbor

[8]Medical Scientist Training Program, University of Michigan, Ann Arbor

[9]Program in Cellular and Molecular Biology, University of Michigan, Ann Arbor

[*]To whom correspondence should be addressed; E-mail: indikar@umich.edu.

Abstract

Every human somatic cell inherits a maternal and paternal genome, which work together to give rise to cellular phenotype. However, the allele-specific relationship between gene expression and genome structure through the cell cycle is largely unknown. By integrating haplotype-resolved genome-wide chromosome conformation capture, mature and nascent mRNA, and protein binding, we investigate this relationship both globally and locally. We introduce the maternal and paternal 4DN, enabling detailed analysis of the mechanisms and dynamics of genome structure and gene function for diploid organisms. Our analyses find significant coordination between allelic expression biases and local genome conformation, and notably absent expression bias in universally essential cell cycle and glycolysis genes. We offer a model in which coordinated biallelic expression reflects prioritized preservation of essential gene sets.

Introduction

The redundancy inherent in the diploid genome makes it robust to potentially harmful mutations. Dependence on a single allele, for example in genes with monoallelic or allele biased expression (MAE, ABE), may increase vulnerability to disease. In familial cancer syndromes, individuals who inherit genetic loss-of-function mutations have a single functional allele, referred to as loss of heterozygosity (LOH) (*1*). BRCA1 and BRCA2 are quintessential examples, where LOH leads to a significant increase in the likelihood of developing breast cancer (*2, 3*). Imprinted genes are also associated with multiple disease phenotypes such as Angelman and Prader-Willi syndromes (*4, 5*). This raises the possibility that other MAE or ABE genes are be associated with disease, but the contribution of allelic bias to disease phenotypes remains poorly understood.

A major step towards understanding the contribution of allelic bias to disease will be to identify ABE genes and gene modules, recognizing that functionally important biases may be transient and challenging to detect. Differences in allelic expression may result from single nucleotide variants (SNVs), insertions, deletions (InDels), and chromatin modifications (*6–10*). By some estimates, ABE occurs in 4-26% of expressed genes across tissues and individuals (*11*). In addition, higher order chromatin conformation and spatial positioning in the nucleus shape gene expression (*12–15*). This may lead to allele-specific differences as the maternal and paternal alleles have been shown to be distant in the nucleus through techniques such as fluorescent in situ hybridization (FISH) (*16, 17*). Allele-specific expression and 3D structures are frequently not distinguished in genomics methods such as RNA-sequencing and Hi-C (genome-wide chromosome conformation capture), complicating interpretation of structure-function relationships. However, complete phasing of the two genomes remains a significant challenge in the field.

To uncover ABE and its relationship to genome structure over time, we integrated allele-specific RNA-seq, Bru-seq, and Hi-C across three phases of the cell cycle in B-lymphocytes (GM12878, **Figure 1**). The RNA-seq and Bru-seq data were separated into their maternal

and paternal components through SNVs/InDels. The Hi-C data was separated using our novel phasing algorithm, HaploHiC, which utilizes SNVs/InDels to impute Hi-C reads of unknown parental origin. We incorporated publicly available allele-specific protein binding data (ChIP-seq) to better understand the mechanisms behind the gene expression and genome structure relationship (*7, 18*). We performed a genome-wide investigation for gene expression biases, chromatin compartment switching, and topologically associated domain (TAD) boundary changes between alleles and cell cycle phases. Along with known XCI and imprinted genes, we found notable patterns of ABE across the cell cycle for hundreds of genes that are not well appreciated in current literature, and corresponding allelic biases in protein binding at many of these loci. We found that differentially expressed genes were significantly more likely to have changes in their local genomic structure than random allele-specific genes. Additionally, we observed a pronounced lack of ABE in crucial biological pathways and essential genes. Our integrated analysis represents an allele-specific extension of the 4D Nucleome (4DN), which describes the relationship between genome structure, gene expression, and cell phenotype through time (*19–21*). Taken together, our findings demonstrate the value of analyzing ABE in the context of corresponding structural and epigenetic alterations across the cell cycle. This approach will be critical in future investigations of human phenotypic traits and their penetrance, genetic diseases, vulnerability to complex disorders, and tumorigenesis.

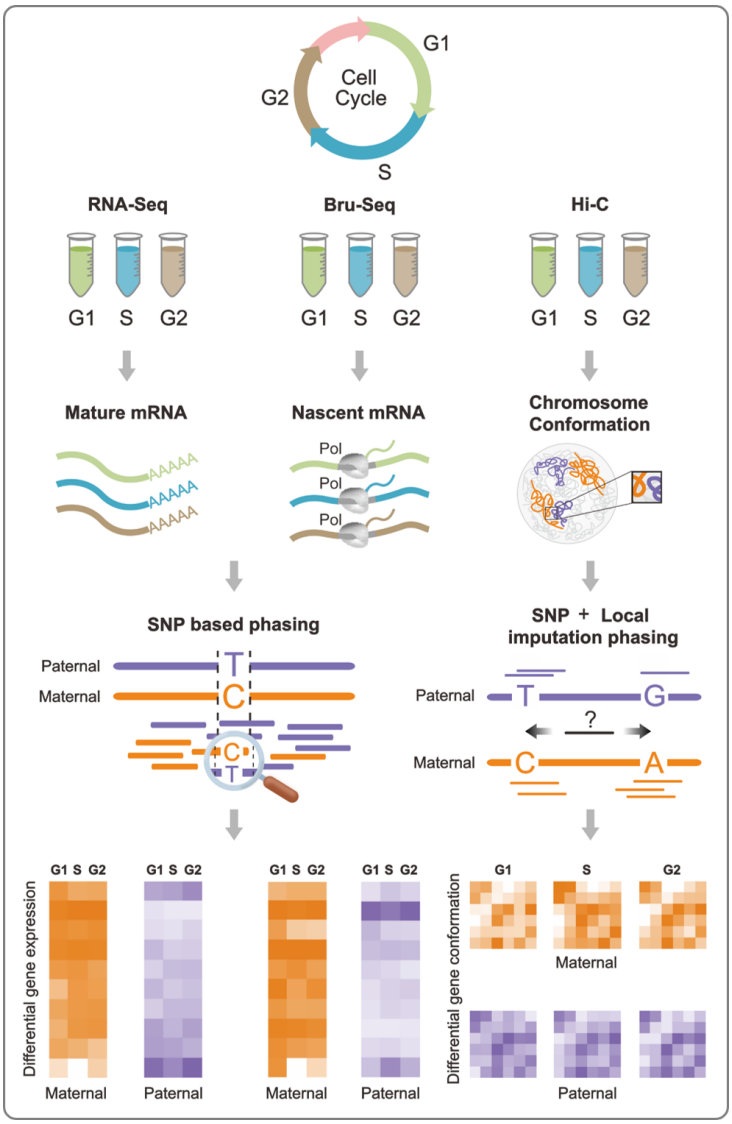


Fig. 1: Experimental and allelic separation workflow. Cell cycle sorted cells were extracted for RNA-seq, Bru-seq, and Hi-C (left to right, respectively). RNA-seq and Bru-seq data were allelically phased via SNVs/InDels (left). SNV/InDel based imputation and haplotype phasing of Hi-C data using HaploHiC (right). These data provide quantitative measures of structure and function of the maternal and paternal genomes through the cell cycle.

Results

Chromosome-Scale Maternal and Paternal Differences

Spatial positioning of genes within the nucleus is known to be associated with transcriptional status (*12–15*). One might expect that the maternal and paternal copies of each chromosome would stay close together to ensure that their respective alleles have equal opportunities for transcription. Imaging of chromosome territories has shown that this is often not the case (**Figure 2A**, **Supplementary Notes C**) (*22–24*). This inspired us to investigate whether the two genomes operate in a symmetric fashion, or if allele-specific differences exist between the genomes regarding their respective chromatin organization patterns and gene expression profiles. We analyzed parentally phased whole-chromosome Hi-C and RNA-seq data at 1 Mb resolution to identify allele-specific differences in both chromosome structure and function, respectively. We subtracted each chromosome's paternal Hi-C matrix from the maternal matrix and found the Frobenius norm of the resulting difference matrix. The Frobenius norm here provides a measure of distance between matrices, where equivalent parental genome structures would result in a value of zero. Similarly, we subtracted the phased RNA-seq vectors in log_2 scale and found the Frobenius norm of each difference vector. The Frobenius norms were adjusted for chromosome size and normalized for both Hi-C and RNA-seq.

We found that all chromosomes have allelic differences in both structure (Hi-C, blue) and function (RNA-seq, red) (**Figure 2B**). Chromosome X had the largest structural difference, as expected, followed by Chromosomes 9, 21, and 14. Chromosome X had the most extreme functional difference as well, followed by Chromosomes 13, 7, and 9. A threshold was assigned at the median Frobenius norm for Hi-C and RNA-seq (**Figure 2B** green dashed lines). The majority of chromosomes with larger differences than the median for Hi-C are also larger than the median in RNA-seq. We found that there is a positive correlation between chromosome level differences in structure and function ($R = 0.66$ and $p < 0.05$).

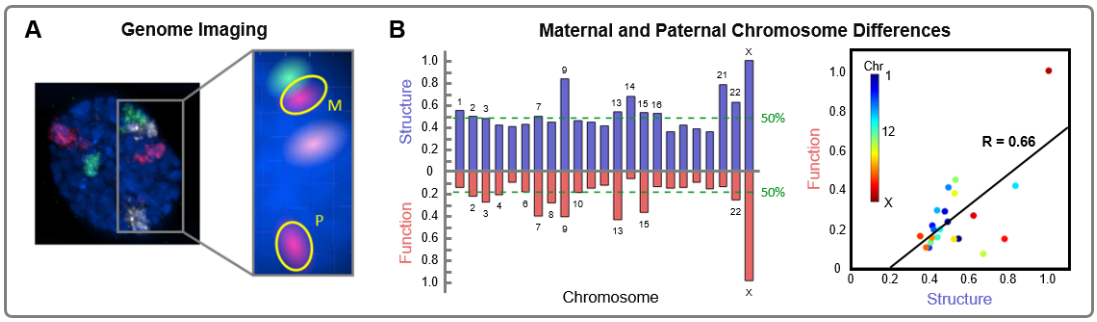


Fig. 2: Genome imaging and chromosome differences. **(A)** Nucleus of a primary human fibroblast imaged using 3D FISH with the maternal and paternal copies of Chromosome 6, 8, and 11 painted red, green, and white, respectively (left). Subsection highlighting the separation between the maternal and paternal copies of Chromosome 11, now colored red (right). **(B)** Chromosome level structural and functional parental differences of GM12878 cells. Structural differences (blue) represent the aggregate changes between maternal and paternal Hi-C over all 1 Mb loci for each chromosome, adjusted for chromosome size in G1. Functional differences (red) represent the aggregate changes between maternal and paternal RNA-seq over all 1 Mb loci for each chromosome, adjusted for chromosome size in G1. Green dashed lines correspond to the median normalized Frobenius norms of structural (0.48, chromosome 3) and functional (0.20, chromosome 6) differences, in the top and bottom respectively, and all chromosomes equal to or greater than the threshold are labeled. Scatter plot of maternal and paternal differences in structure and function with best-fit line ($R = 0.66$ and $p < .05$).

Allele-Specific RNA Expression

After confirming allelic differences in RNA expression at the chromosomal scale, we examined allele-specific expression of individual genes through RNA-seq and Bru-seq. We hypothesized that the chromosome scale expression differences were not only caused by known cases of ABE such as XCI and imprinting, but by widespread ABE over many genes. Therefore, we evaluated all allele-specific genes (genes with informative SNVs/InDels) for differential expression between the six settings: maternal and paternal in G1, S, and G2/M (G2/M referred to as G2 for brevity). These settings give rise to seven comparisons which consist of maternal versus paternal within each of the cell cycle phases (three comparisons), as well as G1 versus S and S versus G2 for the maternal and paternal genomes, respectively (two comparisons for each genome). Differential expression between alleles or cell cycle phases was based on an adjusted

p-value (p_{adj}) < 0.05 from the MATLAB implementation of DESeq, and a fold change (FC) $>$ 2 (**Supplementary Notes A**) (*25*).

First, we identified genes with ABE and allele-specific cell cycle-biased expression (CBE) from RNA-seq. From 23,277 RefSeq genes interrogated, there were 4,727 genes with sufficient coverage on SNVs/InDels to reliably determine allele-specific expression (*26*). We performed differential expression analysis for the seven comparisons to identify which of the 4,727 genes had ABE or CBE (*25*). We identified 778 differentially expressed genes from RNA-seq: 605 ABE genes, 284 CBE genes, and 111 genes with both ABE and CBE (**Supplementary Table 2, 4**). Both exons and introns containing informative SNVs/InDels were used for our Bru-seq data, from which 5,830 allele-specific genes had sufficient coverage (**Supplementary Notes A**). We identified 556 differentially expressed genes from Bru-seq: 438 ABE genes, 171 CBE genes, and 53 genes with both ABE and CBE (**Supplementary Table 3, 5**). We then separated the differentially expressed genes into their respective chromosomes to observe their distribution throughout the genome. From RNA-seq (Bru-seq), we found that autosomes had 5-16% ABE (3-10%) in their allele-specific genes which is similar to previous findings (*11*). As expected, the X Chromosome had a particularly high percentage of ABE genes at 91% (89%).

We identified 316 genes that have ABE in all cell cycle phases from RNA-seq (177 up-regulated in paternal, 139 in maternal) and 175 from Bru-seq (130 upregulated in paternal, 45 in maternal). This is the most common differential expression pattern among ABE genes and these genes form the largest clusters in **Figure 3A**. These clusters include, but are not limited to, XCI, imprinted, and other MAE genes. Known examples within these clusters are highlighted in the 'X-Linked' and 'Imprinted' sections of **Figure 3B**. We also identified hundreds of ABE genes that are not currently appreciated in literature, including genes like *TUBB6* and *NPNT*, shown at the top of the 'Autosomal Genes' section of **Figure 3B**. These genes were strongly biased toward the maternal allele in all cell cycle phases and appeared to be transcriptionally repressed due to their maternally biased nascent transcription. Approximately half of all ABE genes were only differentially expressed in one or two cell cycle phases, leading to the many

smaller clusters seen in **Figure 3A**. Four examples of genes with this behavior are presented in the 'Autosomal Genes' section of **Figure 3B**. Transient expression biases like these may be due to coordinated expression of the two alleles in only certain cell cycle phases, though a mechanism is unclear.

Among the ABE genes from RNA-seq analysis, we found 117 MAE genes. In addition to the requirements for differential expression, we impose the thresholds of a FC $\geq$ 10 and for the inactive allele to have $<$ 0.1 Fragments Per Kilobase of transcript per Million (FPKM), or FC $\geq$ 50 across all cell cycle phases. Our analysis confirmed MAE for imprinted and XCI genes, examples shown in **Figure 3B**. Imprinted and XCI genes are silenced via transcriptional regulation, which was verified by monoallelic nascent RNA expression (Bru-seq). The *XIST* gene, which is responsible for XCI, was expressed in the maternal allele reflecting the deactivation of the maternal X Chromosome. XCI was also observed from Hi-C through large heterochromatic domains in the maternal X Chromosome, and the absence of these domains in the paternal X Chromosome (**Figure 4C**). The inactive X Chromosome in our cells is opposite of what is commonly seen for the GM12878 cell line in literature (likely due to our specific GM12878 sub-clone), but is consistent between our data modalities (*9, 10, 18*). The MAE genes also include six known imprinted genes, four expressed from the paternal allele (*KCNQ1OT1*, *SNRPN*, *SNURF*, and *PEG10*) and two from the maternal allele (*NLRP2* and *HOXB2*). Some of the known imprinted genes that were confirmed in our data are associated with imprinting diseases, such as Beckwith-Wiedemann syndrome (*KCNQ1OT1*), Angelman syndrome (*SNRPN* and *SNURF*), and Prader-Willi syndrome (*SNRPN* and *SNURF*) (*27, 28*). These genes and their related diseases offer further support for allele-specific analysis, as their monoallelic expression could not be detected otherwise.

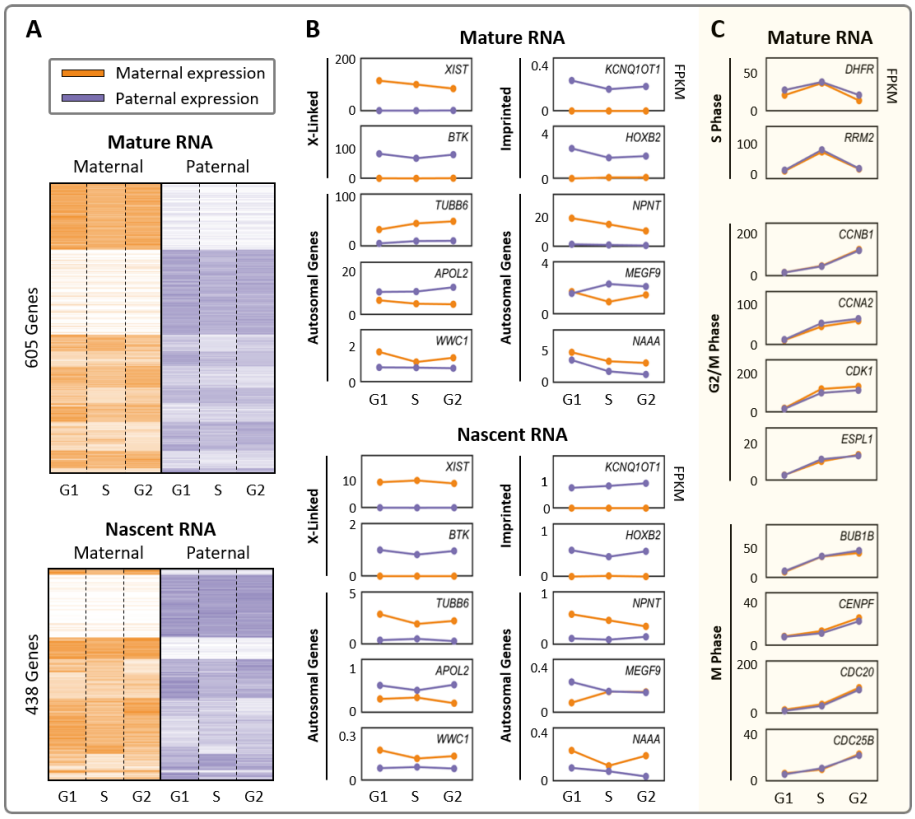


Fig. 3: Allele-specific mature and nascent RNA expression. **(A)** Differentially expressed genes' maternal and paternal RNA expression through the cell cycle. Expression heatmaps are row normalized and clustered by differential expression patterns (**Supplementary Figure 1**). **(B)** Representative examples of X-linked, imprinted, and other genes with allelic bias. Top shows mature RNA levels (RNA-seq) and bottom shows nascent RNA expression (Bru-seq) in both (A) and (B). **(C)** Examples of cell cycle regulatory genes' mature RNA levels through the cell cycle. These genes are grouped by their function in relation to the cell cycle and all exhibit CBE.

After observing that approximately half of all ABE genes had transient expression biases, we hypothesized that alleles may have unique dynamics through the cell cycle. We then focused our investigation on allele-specific gene expression through the cell cycle (**Supplementary Figure 1B**). We compared the expression of each allele between G1 and S as well as between S and G2, which provides insight into the differences between maternal and paternal alleles' dynamics across the cell cycle. In the G1 to S comparison, there are 93 (53) genes in RNA-seq (Bru-seq) which have similar expression dynamics in both alleles. These genes' maternal and paternal

alleles are similarly upregulated or downregulated from G1 to S. In contrast, 124 (104) genes in RNA-seq (Bru-seq) have different expression dynamics between alleles. That is, only one allele is up or downregulated in the transition from G1 to S. In the S to G2 comparison, there are 27 (4) genes in RNA-seq (Bru-seq) that have similar expression dynamics in both alleles and 74 (15) genes with different expression dynamics between alleles. From these data, we see a coordination of expression between many, but certainly not all, alleles through the cell cycle.

Biallelic Expression and Cellular Function

We observed from our analysis of CBE that multiple cell cycle regulatory genes had no instances of ABE (**Figure 3C**). We expanded this set of genes, or biological module, to include all allele-specific genes contained in the KEGG cell cycle pathway (*29*). Again, we observed zero instances of ABE. We hypothesized that genes with crucial cellular functions, like cell cycle regulation, may have coordinated biallelic expression to ensure their sufficient presence. This was further motivated by previous findings which showed restricted genetic variation of enzymes in the essential glycolytic pathway (*30*).

We tested additional modules derived from KEGG pathways containing at least five allele-specific genes, with the circadian rhythm module supplemented by a known core circadian gene set (*19*). Examples of modules with varying proportions of ABE are shown in **Table 1** (**Supplementary Table 7**), where "Percent ABE" refers to the proportion of genes with ABE to the total number of allele-specific genes in that module. We found that there are multiple crucial modules, including the glycolysis/gluconeogenesis and pentose phosphate pathways, which also had zero ABE genes. To explore the possibility of a global phenomenon by which genes essential to cellular fitness are significantly less likely to have biased expression, we analyzed the frequency of ABE in 1,734 genes experimentally determined to be essential in human cells (*31*). Using the 698 allele-specific genes out of this set, we found that these essential genes were significantly less likely to have ABE than a random selection of allele-specific genes (6.0% versus 13.0%, $p < 0.001$, **Supplementary Notes G**), consistent with our hypothesis that critical

genes are likely to be expressed by both alleles. In total, we offer a model in which coordinated biallelic expression reflects prioritized preservation of essential gene sets.

Table 1: Allelic bias in biological modules.

Module	Percent ABE
Cell Cycle	0%
Glycolysis/Gluconeogenesis	0%
Pentose Phosphate	0%
BCR Signaling	7%
p53 Signaling	10%
Circadian Rhythm	17%
Wnt Signaling	19%
Hippo Signaling	23%
Whole Genome	**13%**
Pathways with 0% ABE are highlighted	

Phasing by HaploHiC

To examine the 3D structure of chromatin, we use high-throughput genome-wide chromatin conformation capture, or Hi-C (*9, 32*). Hi-C allows us to create a matrix that can represent all of the contacts between loci in a given region (whole genome, specific chromosome, etc). In this study, we separate the maternal and paternal genomes' contributions to this contact matrix to analyze their similarities and differences in genome structure. In order to determine which Hi-C reads come from which parental origin, we utilize differences in genomic sequence at SNVs/InDels. As these variations are unique to the maternal and paternal genomes, they can be used to quickly and easily separate reads that cover them.

When attempting to separate the maternal and paternal genomes, complications arise when there are sections of DNA that are identical. Due to the cells' ability to correct errors that arise in their DNA, there are a relatively small number of variations. This means our segregated maternal and paternal contact matrices will be extremely sparse if we only separate the small subset of genomic loci with SNVs/InDels. In order to combat this problem, we look to reintroduce contacts of unknown parental origin.

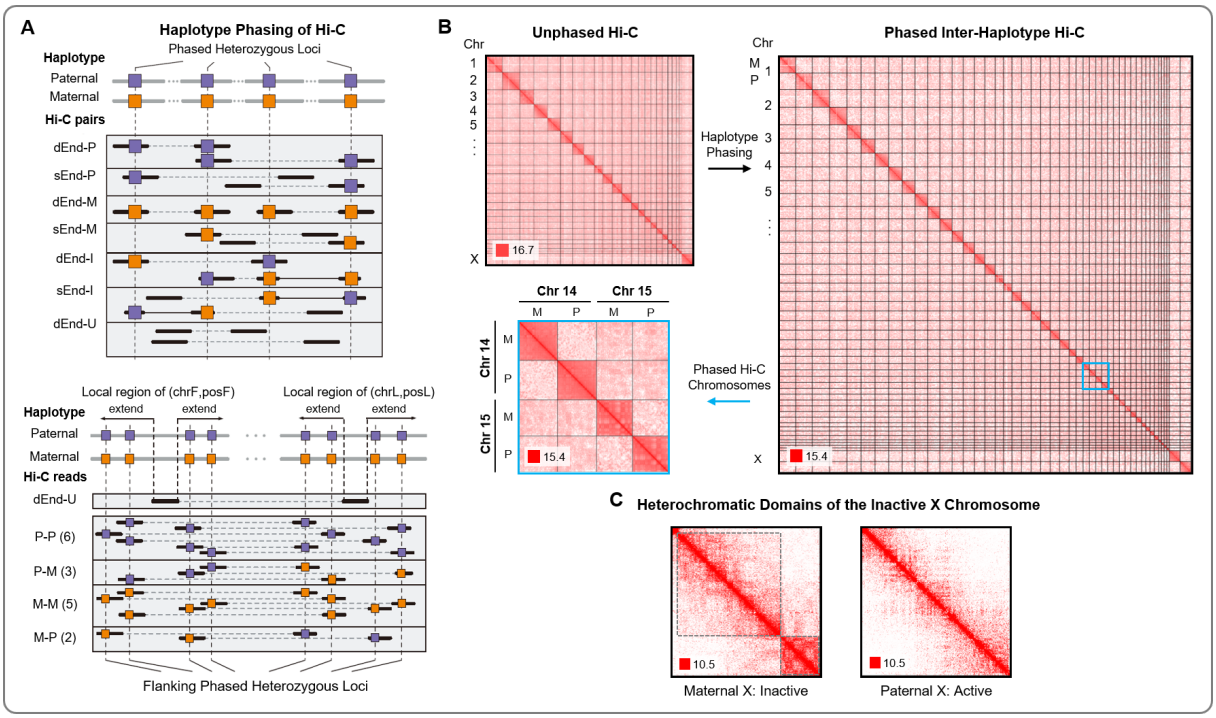


Fig. 4: Haplotype phasing of Hi-C data. **(A)** HaploHiC separates paired-end reads into groups based on parental origin determined through SNVs/InDels (left, **Supplementary Notes B.4**). Reads are grouped by: (i) reads with one (sEnd-P/M) or both ends (dEnd-P/M) mapped to a single parent, (ii) reads are inter-haplotype, with ends mapped to both parents (d/sEnd-I), and (iii) reads with neither end mapped to a specific parent (dEnd-U) . An example of a paired-end read (dEnd-U) with no SNVs/InDels has its origin imputed using nearby reads (right, **Supplementary Notes B.7**). A ratio of paternally and maternally mapped reads is found in a dynamically sized flanking region around the haplotype-unknown read's location (**Supplementary Notes B.5**). This ratio then determines the likelihood of the haplotype-unknown read's origin. **(B)** Whole-genome Hi-C of GM12878 cells (top left). Inter- and intra-haplotype chromatin contacts after phasing Hi-C data using HaploHiC (right). Chromosomes 14 and 15 highlight inter- and intra-chromosome contacts within and between genomes (bottom left). Visualized in log_2 scale 1 Mb resolution in G1. **(C)** Haplotype phasing illustrates that the inactive maternal X Chromosome is partitioned into large heterochromatic domains, shown in log_2 scale 100 kb resolution in G1.

Here we propose HaploHiC, a novel technique for phasing reads of unknown parental origin using a local imputation from known reads. Unlike previous haplotyping methods, HaploHiC uses a data-derived ratio based on the hypothesis that if the maternal and paternal genomes

have different 3D structures, we can use the reads with known origin (at SNV/InDel loci) to predict the origin of neighboring unknown reads (**Figure 4A, Supplementary Notes B**) (*33*). For example, if we observe that many contacts between two loci can be directly mapped to the paternal genome but few to the maternal genome, then unphased contacts between those loci are more likely to be from the paternal genome as well. Evaluation in simulation data shows that HaploHiC performs well, with a 97.66% accuracy of predicting unknown origin reads (**Supplementary Table 11**). The allelically separated genomes can be seen in **Figure 4B**. HaploHiC has been made available through a GitHub repository (https://github.com/Nobel-Justin/HaploHiC).

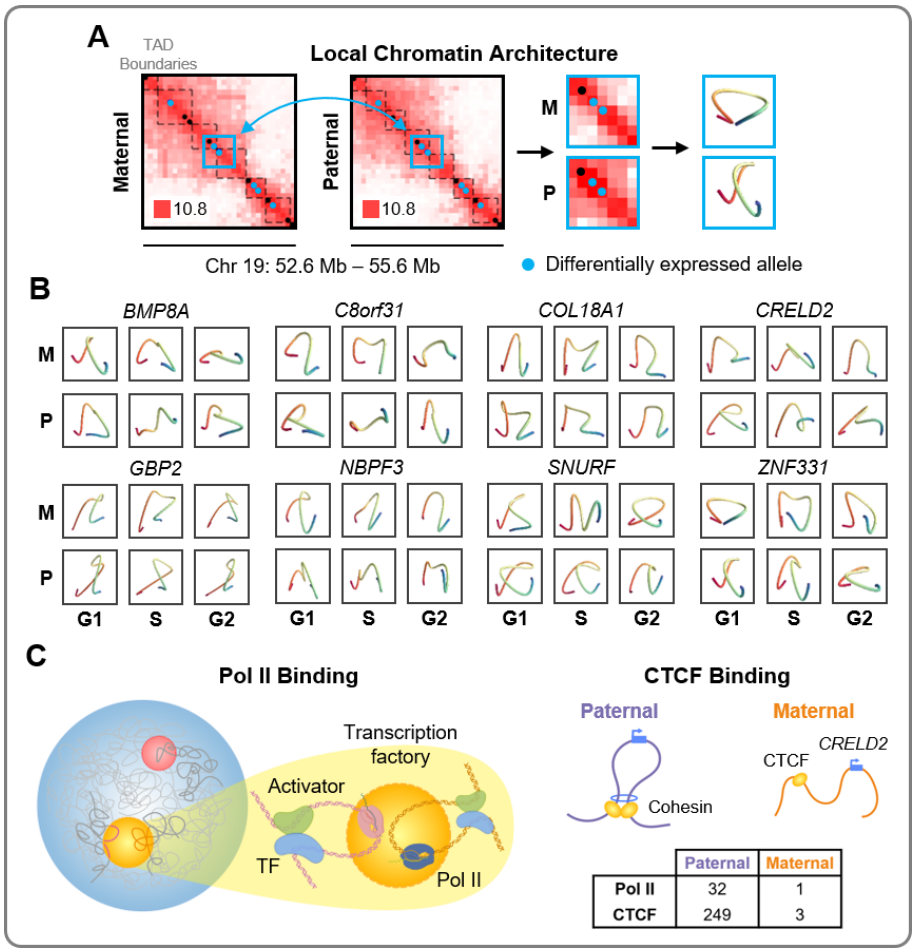
Allele-Specific Genome Structure

Motivated by our observations of chromosome level structural differences between the maternal and paternal genomes, we examined the HaploHiC separated data in more detail to determine where these differences reside. The genome is often categorized into two compartments: transcriptionally active euchromatin and repressed heterochromatin. In studies comparing multiple types of cells or cells undergoing differentiation, areas of euchromatin and heterochromatin often switch corresponding to genes that are activated/deactivated for the specific cell type (*34*). We explored this phenomenon in the context of the maternal and paternal Hi-C matrices to determine if the two genomes had differing chromatin compartments. Chromatin compartments can be identified from Hi-C data using methods such as principal component analysis or spectral clustering (**Supplementary Notes D**) (*35*). We applied spectral clustering to every chromosome across all cell cycle phases at 1 Mb resolution. We found that there were slight changes in chromatin compartments for all chromosomes, but the vast majority of these changes took place on the borders between compartments rather than an entire region switching compartments. These border differences were not enriched for ABE genes. This implies that, although the structures are not exactly the same, the maternal and paternal genomes have similar overall compartmentalization (aside from the X Chromosome).

We next applied spectral clustering recursively to the Hi-C data at 100 kb resolution to determine whether there were differences in TADs between the two genomes throughout the cell cycle (*35*). While the current understanding of genomic structure dictates that TAD boundaries are invariant (between alleles, cell types, etc), it is also known that "intra-TAD" structures are highly variable (*34, 36, 37*). The spectral identification method has an increased ability to discern these subtle structural changes. We found that TAD boundaries were variable between the maternal and paternal genomes and across cell cycle phases in all chromosomes. This supports previous findings of allelic differences in TADs for single cells, and we predict that they are even more variable across cell types (*35, 37*). Differences in TAD boundaries were observed surrounding MAE genes, ABE genes, and genes with coordinated biallelic expression (**Supplementary Figure 6**). This indicated that changes in TAD boundaries were not directly related to allelic expression differences.

Although we did not find a significant relationship between TAD boundary differences and ABE genes, we observed during this analysis that the local genome structure around the six imprinted genes had noticeable differences. We then sought to analyze all genes with ABE or CBE to find out if they had corresponding structural differences at a local level. We analyzed the local Hi-C matrices for each of the 778 RNA-seq and 556 Bru-seq differentially expressed genes. Using a 300 kb flanking region centered at the 100 kb bin containing the transcription start site (TSS), we isolated a 7x7 matrix (700 kb) for each differentially expressed gene (**Figure 5A, B**). These matrices represent the local genomic structure of the differentially expressed genes, and are slightly smaller than average TAD size. We then compared the correlation matrices of the log_2-transformed local Hi-C data and determined whether or not the matrices have statistically significant differences ($p < 0.05$) (*38, 39*). We applied this comparison to all genes that were differentially expressed in RNA-seq (Bru-seq) and found that 627 (446) genes had at least one comparison in which both the expression and local structure had significantly changed. While changes in local genome structure and changes in gene expression do not have a one-to-one relationship, we found that both ABE and CBE genes are more likely to have significant archi-

tectural differences than randomly sampled allele-specific genes ($p < 0.05$) (**Supplementary Figure 7 and Notes G**). This lends further support to the idea that there is a relationship between allele-specific differences in gene expression and genome structure.



	Paternal	Maternal
Pol II	32	1
CTCF	249	3

Fig. 5: Local chromatin structure and transcription factor binding. **(A)** Local regions around DE genes are tested for significant conformation changes. These regions are modeled to visualize the conformations around each allele through G1, S, and G2 (**Supplementary Notes F**). Example of local chromatin structure extraction is shown for *ZNF331* in G1 phase. Hi-C matrices are shown in log_2 scale 100 kb resolution. **(B)** 3D models of the local genome structure around 8 ABE genes with bias agreement of Pol II and significant changes in local genome structure. **(C)** Schematic representation of allele-specific Pol II and CTCF binding, with highlighted gene *CRELD2* which had binding biases in both. Table shows binding biases of Pol II and CTCF on *CRELD2* as an example.

Allele-Specific Protein Binding

To uncover the mechanisms behind the relationship between allele-specific gene expression and genome structure, we looked to DNA binding proteins such as RNA polymerase II (Pol II), CCCTC-binding factor (CTCF), and 35 additional transcription factors. We used publicly available protein binding data from AlleleDB in tandem with RNA-seq and found 120 genes that have an allelic bias in both gene expression and binding of at least one such protein (*18*). We identified 18 genes which have ABE and biased binding of Pol II, with bias agreement in 13 cases (72.2%). That is, the allelic expression and Pol II binding were biased toward the same allele. For CTCF, 45 of 92 cases have bias agreement (48.9%) and for all other transcription factors analyzed, 20 of 30 cases have bias agreement (66.7%) (**Supplementary Table 6**). The CTCF binding bias agreement of around 50% is expected, based on previous studies (*7*). This is likely due to CTCF's role as an insulator, since an allele could be expressed or suppressed by CTCF's presence depending on the context. To avoid potential inconsistencies between our data and the protein data from AlleleDB, we excluded the X Chromosome when testing for ABE and protein binding biases.

We analyzed the 13 genes with allelic expression and Pol II binding bias agreement further to determine if they also had significant changes in local genome structure. Through local Hi-C comparisons, we found that 11 of these genes had significant changes in structure in at least one cell cycle phase. 3D models for 8 of these 11 genes are shown in **Figure 5B**, which highlight differences in local genome structure (**Supplementary Notes F**) (*40*). The genes with bias agreement and changes in local genome structure include known imprinted genes such as *SNURF* and *SNRPN*, as well as genes with known allele-specific expression (and suggested imprinting in other cell types) like *ZNF331* (*41*). Additionally, there are multiple genes with known associations with diseases or disorders such as *BMP8A*, *COL18A1*, *CRELD2*, and *NBPF3* (*42–45*). These findings suggest that changes in local structure often coincide with changes in expression due to the increased or decreased ability of a gene to access the necessary transcriptional machinery within transcription factories. We visualize this relationship for the

gene *CRELD2* as an example (**Figure 5C**) (*14*).

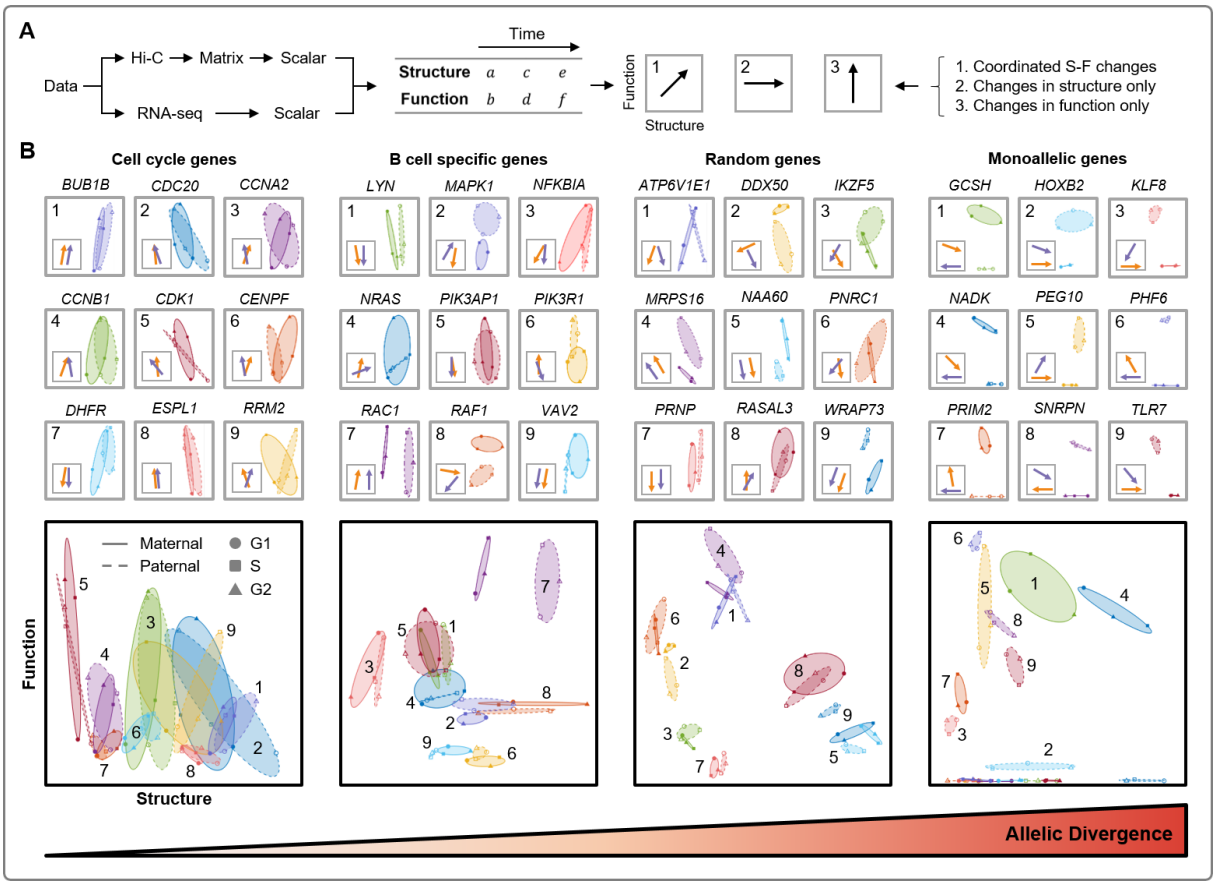


Fig. 6: 4DN phase planes reveal a wide range of allelic divergences in gene sub-networks. **(A)** Workflow to obtain structure and function measures. Simplified phase planes are shown with linear relationships between changes in structure and function, changes in structure with no changes in function, and changes in function with no changes in structure. **(B)** 4DN phase planes of cell cycle genes, genes specific to B cell function, random allele-specific genes, and MAE genes, highlighting the similarities and differences between their alleles. Genes such as *BUB1B* and *PIK3AP1* have similar phase planes between alleles, while *RAC1* differs in structure and *WRAP73* differs in function. The bottom plot for each column combines the phase planes of the nine example genes.

The Maternal and Paternal 4D Nucleome

We define the maternal and paternal 4DN as the integration of allele-specific genome structure with gene expression data through time. Many complex dynamical systems are investigated using a network perspective (*46, 47*). A network is a simplified representation of a system, which captures patterns of interactions between its components and how those interactions change over time. We can consider genome structure as a network, since Hi-C data captures interactions between genomic loci (*12, 15*). In network science, the relative importance of a node in a network is commonly determined using network centrality (*46*). For Hi-C data, we treat genomic loci as nodes and use network centrality to measure the importance of each locus (*39,48*). We performed a global analysis of the maternal and paternal 4DN by combining RNA-seq with multiple network centrality measures. We found differences between the maternal and paternal genomes and across cell cycle phases, but only the X Chromosome had clear separation of the maternal and paternal copies (**Supplementary Figure 9, 10**).

In our earlier analysis we found a significant relationship between ABE and changes in local genome structure. We also observed that genes in critical biological modules had coordinated biallelic expression. Motivated by these results, we performed a local analysis of structure and function together to determine the allele-specific dynamics. We used the eigenvector centrality to quantify the structure of each sub-network (similar to Google's PageRank) (*49*). We utilize the concept of a phase plane to plot the 4DN (4DN phase plane) (**Figure 6**) (*19*). We designate one axis as a measure of structure and the other as a measure of function. Coordinates of each point in the 4DN phase plane are determined from structure data (x-axis, eigenvector centrality derived from Hi-C) and function data (y-axis, RNA-seq). The 4DN phase plane contains three points for each allele, which represent G1, S, and G2. We define the difference between the maternal and paternal alleles in the 4DN phase plane as the allelic divergence (**Supplementary Notes E**).

We show four phase planes of gene sub-networks with various allelic divergences. Genes which are known to be crucial for cell cycle regulation have a small average allelic divergence

of 0.02 (**Figure 6B**, left). This highlights the robust coordination between the alleles to maintain proper cell cycle function. Given that GM12878 is a B-lymphocyte cell line, we were interested in the allelic divergence of genes which are related to B cell receptor functionality. We found that these genes have a moderate allelic divergence of 0.61 (**Figure 6B**, middle-left). This is slightly smaller than a randomly selected set of allele-specific genes whose average allelic divergence was 0.65 (**Figure 6B**, middle-right). MAE genes have large average allelic divergence of 4.7, as expected (**Figure 6B**, right). While this analysis does not offer a statistical significance for structural and functional differences, it represents a broadly applicable approach for understanding the similarities and differences of the maternal and paternal genomes through time. This analysis can be applied to any cell type with allele-specific time-series or cell cycle sorted data.

Discussion

In this study, we present evidence for the intimate relationship among allele-specific gene expression, genome structure, and protein binding across the cell cycle. We validated our data and methods using known allele-specific properties such as the MAE of imprinted and X-linked genes, broad similarities of chromatin compartments between the maternal and paternal genomes, and large heterochromatic domains of the X Chromosome (*9, 50–53*). Unique to this study, we established a coordination of ABE and changes in local genome structure, which included hundreds of genes not commonly associated with ABE. In addition, we offer mechanistic evidence of this coordination through allele-specific protein binding biases. Through the integration of these structural and functional data through time, we developed an extension of the 4DN for the maternal and paternal genomes.

Through our analysis of mature (nascent) RNA, we found 605 (438) genes to be differentially expressed between the two alleles and 284 (171) genes with differential expression through the cell cycle. Approximately half of the genes with ABE are only differentially expressed in certain cell cycle phases, and over half of the genes with CBE are only differentially

expressed in one allele or the other. Further research is needed to explore why certain genes have coordinated cell cycle dynamics across both alleles, while other genes have disparate expression in some cell cycle phases. We predict that these transient allelic biases may be associated with developmental pathologies and tumorigenesis, similar to imprinted and other MAE genes. Conversely, we found no ABE from genes in multiple biological modules, such as the cell cycle and glycolysis pathways (Table 1). We are not able to establish a statistical significance here due to the limited number of allele-specific genes in these modules, so we surveyed a set of 698 essential genes and found that they are significantly less likely to have ABE. This supports our hypothesis of highly coordinated biallelic expression in universally essential genes.

We developed a novel phasing algorithm, HaploHiC, which uses Hi-C reads mapped to SNVs/InDels to predict nearby reads of unknown parental origin. This allowed us to decrease the sparsity of our allele-specific contact matrices and increase confidence in our analysis of the parental differences in genome structure. While we found that the overall compartmentalization (euchromatin and heterochromatin) of the two genomes was broadly similar, there were many differences in TAD boundaries and local genome structure between the two genomes and between cell cycle phases. We focused our search of allele-specific differences in genome structure by calculating the similarity of local contacts surrounding the TSS of differentially expressed genes (*38, 39*). We found that differentially expressed genes were significantly more likely to have corresponding changes in local genome structure than random allele-specific genes.

We incorporated publicly available allele-specific protein binding data for Pol II and CTCF to explore the mechanisms behind the gene expression and local genome structure relationship (*7*). In genes that had both ABE and Pol II binding biases, we found that 72% of these genes had allelic bias agreement. Additionally, we found that 11 of the 13 genes with bias agreement also had significant changes in local genome structure. Analysis of the relationships between allele-specific gene expression, genome structure, and protein binding is currently limited by the amount of information available. The ability to separate maternal and paternal gene expression

and protein binding is dependent on the presence of a SNV/InDel within the gene body and nearby protein binding motifs. As SNVs/InDels are relatively rare in the human genome, the number of genes available to study is severely lacking. Once we are able to separate the maternal and paternal genomes through advances in experimental techniques, we will be able to fully study these relationships.

Overall, these data support an intimate allele-specific relationship between genome structure and function, coupled through allele-specific protein binding. Changes in genome structure, influenced by the binding of proteins such as CTCF, can affect the ability of transcription factors and transcription machinery to access DNA. This results in changes in the rate of transcription of RNA, captured by Bru-seq. The rate of transcription leads to differential steady state gene expression, captured by RNA-seq. Integration of these data into a comprehensive computational framework led to the development of a maternal and paternal 4DN, which can be visualized using phase planes and quantified using allelic divergence. allele-specific analysis across the cell cycle will be imperative to discern the underlying mechanisms behind many diseases by uncovering potential associations between deleterious mutations and allelic bias, and may have broad translational impact spanning cancer cell biology, complex disorders of growth and development, and precision medicine.

Materials and Methods

Cell Culture and Cell Cycle Sorting

We grew GM12878 cells in RPMI1640 medium supplemented with 10% fetal bovine serum (FBS). Live cells were stained with Hoechst 33342 (Cat #B2261, Sigma-Aldrich), and then sorted by fluorescence-activated cell sorting (FACS) to obtain cell fractions at the corresponding cell cycle phases G1, S, and G2/M (**Supplementary Figure 2**). For cells used in construction of Hi-C libraries, cells were crosslinked with 1% formaldehyde, the reaction was neutralized with 0.125 M glycine, then cells were stained with Hoechst and sorted into G1, S, and G2/M

fractions.

RNA-seq and Bru-seq Library Construction

Total RNA were extracted from sorted live cells for RNA-seq and Bru-seq. RNA-seq and Bru-seq analysis were performed as previously described (*54, 55*). Detailed materials and methods are provided in **Supplemental Notes A**.

Hi-C Library Construction and Phasing

Hi-C library construction and Illumina sequencing was done using established methods (*9*). Haplotype phasing of the Hi-C data was performed using a novel method, HaploHiC. Briefly, HaploHiC marks paired-end reads as haplotype-known or -unknown depending on their coverage of heterozygous SNVs and InDels. Haplotype-known reads are directly assigned to their corresponding haplotype, maternal or paternal. HaploHiC uses a local contacts based algorithm to impute the haplotype of haplotype-unknown reads using nearby SNVs and InDels. If the necessary amount of SNVs and InDels for imputation is not reached, HaploHiC randomly assigns the haplotype-unknown reads to be maternal or paternal. Detailed materials and methods for haplotype phasing and Hi-C construction are provided in **Supplementary Notes B**.

Data Availability

The RNA-seq, Bru-seq, and Hi-C data used for this study are available through the Gene Expression Omnibus (GEO) database, https://www.ncbi.nlm.nih.gov/geo/ (accession number GSE159813).

Acknowledgements

We thank the University of Michigan Sequencing Core members for producing high quality Bru-seq, RNA-seq, and Hi-C data. We thank Michele Paulson and Dr. Mats Ljungman for

generating the Bru-seq sequencing libraries and for helpful discussion. We also thank Dr. Jacob Kitzman for providing the GM12878 cell line and assisting in resolving genotype-phasing issues of the cell line. We thank Sam Dilworth and the iReprogram team for processing and formatting the KEGG pathway data. We are grateful to Dr. Thomas Ried for helpful reading of the manuscript. We thank the Air Force Office of Scientific Research (Award No: FA9550-18-1-0028), and the Forbes Institute research fund for supporting our work.

Author contributions

I.R. conceived and supervised the study. H.C. designed and performed the experiments. S.L., W.J., C.C., X.W., and S.R. performed computational analyses. S.L., A.R. L.M. and I.R. interpreted the data. All authors participated in the discussion of the results. S.L., H.C., A.R., L.M., and I.R. prepared the manuscript with input from all authors.

Competing financial interests

The authors declare no competing financial interests.

Materials & Correspondence

Correspondence and material requests should be addressed to I.R.: indikar@umich.edu.

References

1. Alfred G Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, 1971.

2. Julius Gudmundsson, Gudrun Johannesdottir, Jon Thor Bergthorsson, Adalgeir Arason, Sigurdur Ingvarsson, Valgardur Egilsson, and Rosa Björk Barkardottir. Different tumor types from brca2 carriers show wild-type chromosome deletions on 13q12–q13. *Cancer research*, 55(21):4830–4832, 1995.

3. Kara N Maxwell, Bradley Wubbenhorst, Brandon M Wenz, Daniel De Sloover, John Pluta, Lyndsey Emery, Amanda Barrett, Adam A Kraya, Ioannis N Anastopoulos, Shun Yu, et al. Brca locus-specific loss of heterozygosity in germline brca1 and brca2 carriers. *Nature communications*, 8(1):1–11, 2017.

4. Irina S Zakharova, Alexander I Shevchenko, and Suren M Zakian. Monoallelic gene expression in mammals. *Chromosoma*, 118(3):279–290, 2009.

5. Karin Buiting. Prader–willi syndrome and angelman syndrome. In *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, volume 154, pages 365–376. Wiley Online Library, 2010.

6. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

7. Joel Rozowsky, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, Robert Bjornson, Yong Kong, Naoki Kitabayashi, et al. Alleleseq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology*, 7(1):522, 2011.

8. Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317, 2015.

9. Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A

3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

10. Longzhi Tan, Dong Xing, Chi-Han Chang, Heng Li, and X Sunney Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405):924–928, 2018.

11. Danny Leung, Inkyung Jung, Nisha Rajagopal, Anthony Schmitt, Siddarth Selvaraj, Ah Young Lee, Chia-An Yen, Shin Lin, Yiing Lin, Yunjiang Qiu, et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, 518(7539):350, 2015.

12. Indika Rajapakse and Mark Groudine. On emerging nuclear order. *Journal of Cell Biology*, 192(5):711–721, 2011.

13. Tom Misteli. The inner life of the genome. *Scientific American*, 304(2):66, 2011.

14. Peter R Cook. A model for all genomes: the role of transcription factories. *Journal of molecular biology*, 395(1):1–10, 2010.

15. Tom Misteli. The self-organizing genome: Principles of genome architecture and function. *Cell*, 2020.

16. Brian J Beliveau, Alistair N Boettiger, Maier S Avendaño, Ralf Jungmann, Ruth B McCole, Eric F Joyce, Caroline Kim-Kiselak, Frédéric Bantignies, Chamith Y Fonseka, Jelena Erceg, et al. Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using oligopaint fish probes. *Nature communications*, 6(1):1–13, 2015.

17. Thomas Cremer and Marion Cremer. Chromosome territories. *Cold Spring Harbor perspectives in biology*, 2(3):a003889, 2010.

18. Jieming Chen, Joel Rozowsky, Timur R Galeev, Arif Harmanci, Robert Kitchen, Jason Bedford, Alexej Abyzov, Yong Kong, Lynne Regan, and Mark Gerstein. A uniform survey

of allele-specific binding and expression over 1000-genomes-project individuals. *Nature communications*, 7:11101, 2016.

19. Haiming Chen, Jie Chen, Lindsey A Muir, Scott Ronquist, Walter Meixner, Mats Ljungman, Thomas Ried, Stephen Smale, and Indika Rajapakse. Functional organization of the human 4d nucleome. *Proceedings of the National Academy of Sciences*, 112(26):8002–8007, 2015.

20. Thomas Ried and Indika Rajapakse. The 4d nucleome. *Methods (San Diego, Calif.)*, 123:1, 2017.

21. Job Dekker, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C Oshea, Peter J Park, Bing Ren, et al. The 4d nucleome project. *Nature*, 549(7671):219–226, 2017.

22. Irina Solovei, Antonio Cavallo, Lothar Schermelleh, Françoise Jaunin, Catia Scasselati, Dusan Cmarko, Christoph Cremer, Stanislav Fakan, and Thomas Cremer. Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3d-fish). *Experimental cell research*, 276(1):10–23, 2002.

23. Andreas Bolzer, Gregor Kreth, Irina Solovei, Daniela Koehler, Kaan Saracoglu, Christine Fauth, Stefan Müller, Roland Eils, Christoph Cremer, Michael R Speicher, et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol*, 3(5):e157, 2005.

24. Scott Ronquist, Walter Meixner, Indika Rajapakse, and John Snyder. Insight into dynamic genome imaging: Canonical framework identification and high-throughput analysis. *Methods*, 123:119–127, 2017.

25. Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

26. Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.

27. Yang Cao, Susan S AlHumaidi, Eissa A Faqeih, Beth A Pitel, Patrick Lundquist, and Umut Aypar. A novel deletion of snurf/snrpn exon 1 in a patient with prader-willi-like phenotype. *European journal of medical genetics*, 60(8):416–420, 2017.

28. J Adams. Imprinting and genetic disease: Angelman, prader-willi and beckwith-weidemann syndromes. *Nat Educ*, 1(1), 2008.

29. Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

30. PT Wade Cohen, GS Omenn, AG Motulsky, S-H Chen, and ER Giblett. Restricted variation in the glycolytic enzymes of human brain and erythrocytes. *Nature New Biology*, 241(112):229–233, 1973.

31. Vincent A Blomen, Peter Májek, Lucas T Jae, Johannes W Bigenzahn, Joppe Nieuwenhuis, Jacqueline Staring, Roberto Sacco, Ferdy R van Diemen, Nadine Olk, Alexey Stukalov, et al. Gene essentiality and synthetic lethality in haploid human cells. *Science*, 350(6264):1092–1096, 2015.

32. Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

33. Siddarth Selvaraj, Jesse R Dixon, Vikas Bansal, and Bing Ren. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature biotechnology*, 31(12):1111, 2013.

34. Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, et al. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331, 2015.

35. Jie Chen, Alfred O Hero III, and Indika Rajapakse. Spectral identification of topological domains. *Bioinformatics*, 32(14):2151–2158, 2016.

36. Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376, 2012.

37. Elizabeth H Finn, Gianluca Pegoraro, Hugo B Brandao, Anne-Laure Valton, Marlies E Oomen, Job Dekker, Leonid Mirny, and Tom Misteli. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell*, 176(6):1502–1515, 2019.

38. James A Koziol, Joel E Alexander, Lance O Bauer, Samuel Kuperman, Sandra Morzorati, Sean J O'connor, John Rohrbaugh, Bernice Porjesz, Henri Begleiter, and John Polich. A graphical technique for displaying correlation matrices. *The American Statistician*, 51(4):301–304, 1997.

39. Scott Ronquist, Sijia Liu, Michael Perlman, and Indika Rajapakse. 4dnvestigator: a toolbox for the analysis of time series hi-c and rna-seq data. *bioRxiv*, 2020.

40. Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. A statistical approach for inferring the 3d structure of the genome. *Bioinformatics*, 30(12):i26–i33, 2014.

41. Eyal Ben-David, Shahar Shohat, and Sagiv Shifman. Allelic expression analysis in the brain suggests a role for heterogeneous insults affecting epigenetic processes in autism spectrum disorders. *Human molecular genetics*, 23(15):4111–4124, 2014.

42. Fang-Ju Wu, Ting-Yu Lin, Li-Ying Sung, Wei-Fang Chang, Po-Chih Wu, and Ching-Wei Luo. Bmp8a sustains spermatogenesis by activating both smad1/5/8 and smad2/3 in spermatogonia. *Science signaling*, 10(477), 2017.

43. Maria Rita Passos-Bueno, Oscar T Suzuki, Lucia M Armelin-Correa, Andréa L Sertié, Flavia IV Errera, Kelly Bagatini, Fernando Kok, and Katia RM Leite. Mutations in collagen 18a1 (col18a1) and their relevance to the human phenotype. *Anais da Academia Brasileira de Ciencias*, 78(1):123–131, 2006.

44. Yeawon Kim, Sun-Ji Park, Scott R Manson, Carlos AF Molina, Kendrah Kidd, Heather Thiessen-Philbrook, Rebecca J Perry, Helen Liapis, Stanislav Kmoch, Chirag R Parikh, et al. Elevated urinary creld2 is associated with endoplasmic reticulum stress–mediated kidney disease. *JCI insight*, 2(23), 2017.

45. Joseph Petroziello, Andrew Yamane, Lori Westendorf, Melissa Thompson, Charlotte McDonagh, Charles Cerveny, Che-Leung Law, Alan Wahl, and Paul Carter. Suppression subtractive hybridization and expression profiling identifies a unique set of genes overexpressed in non-small-cell lung cancer. *Oncogene*, 23(46):7734–7745, 2004.

46. Mark Newman. *Networks*. Oxford university press, 2018.

47. Steven H Strogatz. Exploring complex networks. *nature*, 410(6825):268–276, 2001.

48. Sijia Liu, Haiming Chen, Scott Ronquist, Laura Seaman, Nicholas Ceglia, Walter Meixner, Pin-Yu Chen, Gerald Higgins, Pierre Baldi, Steve Smale, et al. Genome architecture mediates transcriptional control of human myogenic reprogramming. *iScience*, 6:232–246, 2018.

49. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

50. Wolf Reik and Jörn Walter. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, 2(1):21, 2001.

51. Tomas Babak, Brian DeVeale, Emily K Tsang, Yiqi Zhou, Xin Li, Kevin S Smith, Kim R Kukurba, Rui Zhang, Jin Billy Li, Derek van der Kooy, et al. Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nature genetics*, 47(5):544, 2015.

52. Yael Baran, Meena Subramaniam, Anne Biton, Taru Tukiainen, Emily K Tsang, Manuel A Rivas, Matti Pirinen, Maria Gutierrez-Arcelus, Kevin S Smith, Kim R Kukurba, et al. The landscape of genomic imprinting across diverse adult human tissues. *Genome research*, 25(7):927–936, 2015.

53. Federico A Santoni, Georgios Stamoulis, Marco Garieri, Emilie Falconnet, Pascale Ribaux, Christelle Borel, and Stylianos E Antonarakis. Detection of imprinted genes by single-cell allele-specific gene expression. *The American Journal of Human Genetics*, 100(3):444–453, 2017.

54. Laura Seaman, Haiming Chen, Markus Brown, Darawalee Wangsa, Geoff Patterson, Jordi Camps, Gilbert S Omenn, Thomas Ried, and Indika Rajapakse. Nucleome analysis reveals structure–function relationships for colon cancer. *Molecular Cancer Research*, 15(7):821–830, 2017.

55. Michelle T Paulsen, Artur Veloso, Jayendra Prasad, Karan Bedi, Emily A Ljungman, Brian Magnuson, Thomas E Wilson, and Mats Ljungman. Use of bru-seq and bruchase-seq for genome-wide assessment of the synthesis and stability of rna. *Methods*, 67(1):45–54, 2014.

56. Thomas D Wu and Serban Nacu. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 2010.

57. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

58. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

59. Neva C Durand, Muhammad S Shamim, Ido Machol, Suhas SP Rao, Miriam H Huntley, Eric S Lander, and Erez Lieberman Aiden. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell systems*, 3(1):95–98, 2016.

60. Philip A Knight and Daniel Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33(3):1029–1047, 2013.

61. Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

62. Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

63. Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

64. Nicolas Servant, Nelle Varoquaux, Bryan R Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. Hic-pro: an optimized and flexible pipeline for hi-c data processing. *Genome biology*, 16(1):259, 2015.

65. Simon A Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, et al. Cosmic: somatic cancer genetics at high-resolution. *Nucleic acids research*, 45(D1):D777–D783, 2016.

66. Matthew Z DeMaere and Aaron E Darling. Sim3c: simulation of hi-c and meta3c proximity ligation sequencing technologies. *GigaScience*, 7(2):gix103, 2017.

67. Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Soc., 1997.

68. Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

69. P. Sun and R. M. Freund. Computation of minimum-volume covering ellipsoids. *Operations Research*, 52(5):690–706, 2004.

70. Mikkel B Stegmann and David Delgado Gomez. A brief introduction to statistical shape analysis. *Informatics and mathematical modelling, Technical University of Denmark, DTU*, 15(11), 2002.

71. Prabhu Ramachandran and Gaël Varoquaux. Mayavi: 3d visualization of scientific data. *Computing in Science & Engineering*, 13(2):40–51, 2011.

72. Alberto Santos, Rasmus Wernersson, and Lars Juhl Jensen. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic acids research*, 43(D1):D1140–D1144, 2015.

Supplementary Notes

A Generation and Haplotype Assignment of RNA-seq and Bru-seq

A.1 Determine mRNA abundance using RNA-seq analysis at G1, S, and G2/M

Live cells were cell cycle sorted and used for total RNA extraction. Subsequently, RNA-seq library construction was carried out in the University of Michigan DNA Sequencing Core facilities (Seq-Core), and sequence reads of 50-base in length were generated on an Illumina HiSeq 2500 station.

A.2 Determine allele-specific nascent transcription using Bru-seq analysis at G1, S, and G2/M

We performed 5-bromouridine (Bru) incorporation in live cells for 30 minutes, and the Bru-labeled cells were then stained on ice with Hoechst 33342 for 30 min before sorting at 4°C to isolate G1, S, and G2/M phase cells. The sorted cells were immediately lysed in TRizol (Cat # 15596026, ThermoFisher) and frozen. To isolate Bru-labeled RNA, DNAse-treated total RNA was incubated with anti-BrdU antibodies conjugated to magnetic beads (*32*). We converted the Bru-labeled transcripts from the different samples into cDNA libraries and they were deep-sequenced at 50-base length on an Illumina HiSeq2500 platform.

A.3 Haplotype assignment and analysis of RNA-seq and Bru-seq

RNA-seq and Bru-seq analysis were performed as previously described (*54, 55*). Briefly, Bru-seq used Tophat (v1.3.2) to align reads without de novo splice junction calling after checking quality with FastQC. A custom gene annotation file was used in which introns are included but preference to overlapping genes is given on the basis of exon locations and stranding where possible (See (*55*) for full details). Similarly for RNA-seq data processing, the raw reads were checked with FastQC (version 0.10.1). Tophat (version 2.0.11) and Bowtie (version 2.1.0.0) were used to align the reads to the reference transcriptome (HG19). Cufflinks (version 2.2.1)

was used for expression quantification, using UCSC hg19.fa and hg19.gtf as the reference genome and transcriptome.

A locally developed R script using CummeRbund was used to format the Cufflinks output. To determine allele-specific transcription and gene expression through Bru-seq and RNA-seq, all reads were aligned using GSNAP, a SNV aware aligner (*56*). HG19 and USCS gene annotations were used for the reference genome and gene annotation, respectively. The gene annotations were used to create the files for mapping to splice sites (used with –s option). Optional inputs to perform SNV aware alignment were included. Specifically, –v was used to include the list of heterozygous SNVs (ftp://platgene_ro@ussd-ftp.illumina.com/2016-1.0/hg19/small_variants/NA12878/NA12878.vcf.gz) and –use-sarray=0 was used to prevent bias against non-reference alleles.

After alignment, the output SAM files were converted to BAM files, sorted and indexed using SAMtools (*57*). SNV alleles were quantified using bam-readcounter (https://github.com/genome/bam-readcount) which was used to count the number of each base that was observed at each of the heterozygous SNV locations. Allele-specificity of each gene was then assessed by combining all of the SNVs in each gene. For RNA-seq, only exonic SNVs were used. For Bru-seq, exonic SNVs were counted first, and then non-exonic SNVs were counted if they were in a genes intronic region. Maternal and paternal abundance of each gene were calculated by multiplying the overall abundance estimate by the fraction of the SNV-covering reads that were maternal and paternal, respectively.

RNA-seq and Bru-seq were binned into 100 kb and 1 Mb bins to match the resolution of the Hi-C data. This was done separately using the maternal and paternal expression estimates by adding the expression of the genes in a bin and when necessary dividing a genes counts according the proportion of the bin in each gene. About 75% of genes could not be assessed for allele-specific expression due to a lack of SNVs in the gene body. When binning RNA-seq and Bru-seq, an assumption of 50% maternal and paternal expression was made for these genes to avoid losing that data.

From the 23,277 Refseq genes interrogated, we identified 6,795 transcripts containing informative heterozygous SNVs or InDels. There were 4,727 genes with allele read counts $\geq$5, which is the minimum number of counts that we use to reliably estimate ABE and CBE. We identified differentially expressed genes between alleles and between cell cycle phases using a MATLAB implementation of DESeq (https://www.mathworks.com/help/bioinfo/ug/identifying-differentially-expressed-genes-from-rna-seq-data.html) (*25*). We imposed a minimum FPKM level of 0.1, false discovery rate adjusted p-value threshold of 0.05, and a fold change cutoff of FC $>$ 2 (*58*). We identified a total of 778 genes (FDR $<$ 0.05) which were differentially expressed. 605 of these genes had ABE and 284 genes had CBE (111 genes in common).

Bru-seq detects nascent transcripts containing both exons and introns. Sequence reads from both exonRNA-Seq: a revolutionary tool for transcriptomicss and introns containing informative SNVs are used to evaluate ABE and CBE. We identified 266,899 informative SNVs from the Bru-seq data, while only 65,676 such SNVs from RNA-seq data. However in the Bru-seq data, many SNVs have read coverage depth that is too low to be able to be statistically evaluated compared to that of the RNA-seq data. We require allelic count $\geq$ 5 to estimate ABE and CBE for nascent transcripts as done for the RNA-seq data. This criterion found that there were similar numbers of informative SNVs (19,394 and 19,998) in the RNA-seq and Bru-seq data, respectively. Inclusion of the intron SNVs makes the total number of informative genes (genes tested for differential expression) to be 5,830. We observed that there were larger variances between samples in the Bru-seq data set than in RNA-seq. Differential expression of Bru-seq data was found in the same manner as was done for RNA-seq. We found that 556 genes were differentially expressed: 438 genes with ABE and 171 genes with CBE (53 genes in common).

B Generation and Haplotype Assignment of Hi-C

B.1 Hi-C library construction

Cross-linked chromatin was digested with the restriction enzyme MboI for 12 hours. The restriction enzyme fragment ends were tagged with biotin-dATP and ligated in situ. After ligation,

the chromatins were de-cross-linked, and DNA was isolated for fragmentation. DNA fragments tagged by biotin-dATP, in the size range of 300-500 bp, were pulled down for sequencing adaptor ligation and polymerase chain reaction (PCR). The PCR products were then sequenced on the Illumina HiSeq2500 platform in the Seq-Core facility.

After haplotype assignment through HaploHiC, Hi-C PE-reads are distributed to intra-haploptye (P-P and M-M) and inter-haplotype (P-M and M-P) (**Supplementary Notes B.2-B.9**). Juicer is applied on intra-haploptye PE-reads, and outputs maternal and paternal contact matrices which are normalized through the Knight-Ruiz method of matrix balancing (*59, 60*). Inter-haplotype contact matrices are generated by HaploHiC (**Supplementary Notes B.8**). Inter-haplotype contacts shown in **Figure 4B** and **Supplementary Figure 8**. Both base pair level and fragment level matrices are constructed. The resolution of base pair level matrices are 1 Mb and 100 kb. Gene-level contacts are converted from fragment level matrices by HaploHiC.

B.2 Phased germline mutations of GM12878

From the AlleleSeq database (version: Jan-7-2017) (*7*), we downloaded the VCF file of sample GM12878 . This VCF file contains phased heterozygous germline mutations (SNV and InDel) with a total of 2.49 million genomic loci (**Supplementary Table 8**). We confirmed that paternal allele is the left side of phased alleles, and the maternal allele is the right side, e.g., '0|1' at HG19 genomic position chr1:2276371. This mutation list was also utilized in the realignment process of GATK.

HaploHiC utilizes heterozygous genotype information to distinguish reads' parental origin as long as the reads cover the heterozygous genomic loci. Considering that short sequence insertions and deletions dramatically influence alignment accuracy, we applied a filtration on heterozygous InDels. All homologous and heterozygous InDel alleles were assigned a minimum distance to the next InDel on the haplotyped genome. This minimum distance considers repeatability of both the mutation sequence and local genomic context. The mutation sequence (i.e., the inserted or deleted sequence) might be repetitive and repeated at its genomic position. For ex-

ample, at HG19 genomic position chr1:2277268, the maternal genome has a deletion ('CACA'). The deleted sequence is 'CA'-unit repetitive (2 times) and also repeated in the adjacent reference context ('CA'-unit repeats 11 times). The minimum distance is the whole repeats' length added by unit repeated time of mutation sequence. So, at genomic position chr1:2277268, the minimum distance of the deletion allele on maternal genome is 28, which means any InDel located within this distance on maternal genome will be excluded from following analysis. In total, 11,614 heterozygous genomic loci harboring InDel alleles were filtered (**Supplementary Table 8**). HaploHiC outputs a list of these filtered genotypes. The minimum distance is also applied in InDel allele judgment on sequencing reads (see B.4).

B.3 Alignment and filtering of Hi-C PE-reads

Illumina adapter sequences and low quality ends were trimmed from raw Hi-C reads by Trimmomatic (*61*). Paired-ends (PE, R1 and R2) were separately aligned to the human reference genome (HG19) using the function 'mem' from BWA (0.7.15) with default parameters (*62*). Then, each alignment BAM file was coordinated, sorted, and realigned by GATK (3.5) (*63*) with the VCF file described in **Supplementary Notes B.2**. All realigned BAM files were sorted by read names before processing in HaploHiC.

HaploHiC loads Hi-C PE-reads' mapping information from paired alignment BAM files, and filters unqualified and invalid pairs in the first step. After exclusion of unqualified and invalid pairs, remaining Hi-C PE-reads are eligible to provide chromatin contacts (**Supplementary Table 9**).

Unqualified PE-reads are filtered out if either or both ends are un-mapped, mapped with a mapping quality less than 20, or mapped to multiple genomic positions. Note that supplementary alignments are reserved, because theoretically, a portion of sequencing reads that cover ligation site could be soft-clipped mapped to two genomic locations by BWA and provide contact information. HaploHiC sets two filtering criteria on primary and supplementary alignments: 1) the difference between their length sum and the reads' length must be less than one fifth of

the reads' length, 2) the overlaps between primary and supplementary mapped parts must be shorter than one third of the shorter one. If the two criteria are not satisfied simultaneously, the supplementary alignment will be marked as 'remove_SP' and discarded.

HaploHiC detects four categories of invalid pairs: 1) dangling ends, 2) self circle, 3) dumped pair forwarded mapped, and 4) dumped pair reversed mapped. The definition of these four categories are the same as HiC-Pro (*64*), while the singleton type is included in the unqualified PE-reads mentioned above. The two ends of an invalid pair must map to same the enzyme fragment or within a close distance (1 kb).

B.4 Parental origin categories of Hi-C PE-reads

HaploHiC assigns a haplotype situation to sequencing reads by checking whether reads cover and support heterozygous alleles of specific parental origin. Note that, in this section, we deal with each alignment of sequencing reads. Four haplotype situations are defined (**Figure 4**): 1) 'P' (only paternal allele), 2) 'M' (only maternal allele), 3) 'I' (both maternal and paternal alleles, i.e., inter-haplotype), and 4) 'U' (no parental allele supported). Clearly, the former three situations ('P', 'M', and 'I') are haplotype-known, and the fourth ('U') is haplotype-unknown. Note that sequencing reads with 'I' support both haplotypes simultaneously, as covering the junction site of the ligated maternal and paternal fragments.

For sequencing reads covering heterozygous genomic positions and matching either parental allele, the bases that match the allele must meet two criteria: 1) base quality is not less than 20, and 2) distance to bilateral edges of this read is not less than 5 bp. Note, if the allele is determined through an InDel, the read-edge distance might use the minimum distance representing the repeatability mentioned above (see section B.2) if the latter is larger. HaploHiC outputs a list recording all InDel alleles that have the repeatability distance. There are a total of 255,961 heterozygous loci having such InDel alleles in sample GM12878 (**Supplementary Table 8**).

According to the haplotype status of the two ends, HaploHiC assigns Hi-C PE-reads to seven categories: dEnd-P/M/I, sEnd-P/M/I, and dEnd-U. Here, 'dEnd' and 'sEnd' represent dual-

ends and single-end, respectively. Common instances of each category are shown in **Figure 4**. Additionally, we also set requirements on distance of alignments (**Supplementary Table 9**).

- dEnd-P: both ends support only paternal alleles, and are not close-aligned
- dEnd-M: both ends support only maternal alleles, and are not close-aligned
- dEnd-I: at least one end supports both maternal and paternal alleles, and the other one is haplotype-known
- sEnd-P: one end supports only paternal allele, the other one has haplotype-unknown alignment
- sEnd-M: one end supports only maternal allele, the other one has haplotype-unknown alignment
- sEnd-I: one end supports both maternal and paternal alleles, the other one has haplotype-unknown alignment
- dEnd-U: both ends are haplotype-unknown.

Clearly, four categories (dEnd-P/M/I and sEnd-I) have confirmed allele-specific contacts, and we call them phased Hi-C PE-reads. Conversely, the other three categories are unphased: PE-reads have one (sEnd-P/M) or two (dEnd-U) haplotype-unknown ends. The next step is to assign the haplotype to these unphased Hi-C PE-reads based on the allele-specific contact information in local regions.

B.5 Local region contacts from phased Hi-C PE-reads

By integrating all phased Hi-C PE-reads, HaploHiC records allele-specific contacts in windowed regions. For example, to record allele-specific contacts of two genomic regions ('no.A' window on 'chrA', and 'no.B' window on 'chrB'), HaploHiC sorts 'chrA' and 'chrB' in ASCII

order, and if ‘chrA’ and ‘chrB’ are the same chromosome, it then sorts ‘no.A’ and ‘no.B’ in ascending order. This dual-sorting process avoids duplicated records and reduces software memory consumption. After sorting, the former region is ‘chrF, no.F’ (here, ‘F’ for former), and the latter region is ‘chrL, no.L’ (here, ‘L’ for latter). HaploHiC keeps a dictionary with key-value pairs. The key is ‘chrF, no.F, chrL, no.L’, and its value is a set of phased Hi-C PE-reads that link these two regions with haplotype combinations (i.e., ‘P-P’, ‘M-M’, ‘P-M’, and ‘M-P’, **Figure 4B, Supplementary Figure 4**). Additionally, HaploHiC de-duplicates the Hi-C PE-reads under each haplotype combination.

HaploHiC utilizes a local allele-specific contacts based algorithm to impute the haplotype for haplotype-unknown ends. To get the local region of one end of the unphased Hi-C PE-reads, HaploHiC extends bilaterally from the mapped position. The extension length is 50 kb on each unilateral side, which is referred to as an extension unit. Note that each unilateral region cannot harbor more than 30 phased heterozygous loci, or else the unilateral region will be trimmed. The bilateral extended regions are merged as one local region. A similar extension gets the local region of the other end of the Hi-C PE-read. Then, from the dictionary of phased Hi-C pairs, HaploHiC counts PE-reads linking these two local regions of each haplotype combination. For simplicity, the contact counts of haplotype combinations are: α (‘P-P’), β (‘M-M’), γ (‘P-M’), and δ (‘M-P’), respectively. HaploHiC uses these counts to impute the haplotype for haplotype-unknown reads. Note that if the sum of these counts is zero, local regions of both ends will be iteratively extended by more extension units until the sum is non-zero or local regions reach the maximum length (1 Mb). Finally, if the sum is still zero, we define this pair of local regions as unphased.

B.6 Assign haplotype to unphased Hi-C PE-reads (sEnd-P/M)

First, HaploHiC deals with unphased Hi-C PE-reads from sEnd-P/M categories. Because these Hi-C pairs already have one haplotype-known end, some haplotype combinations should be excluded. For example, one Hi-C pair has one paternal end (mapped to ‘posA’ on ‘chrA’)

and one haplotype-unknown end (mapped to 'posB' on 'chrB'). After dual-sorting the mapped chromosomes and positions, 'chrB, posB' is the 'chrF, posF', and 'chrA, posA' is the 'chrL, posL'. HaploHiC calculates the local regions of 'chrF, posF' and 'chrL, posL' respectively, and summarizes contacts counts of haplotype combinations recorded under key 'chrF, posF, chrL, no.L': α ('P-P'), β ('M-M'), γ ('P-M'), and δ ('M-P'). Because 'chrL, posL' is from the paternal genome in this instance, β ('M-M') and γ ('P-M') are impossible and should be excluded. HaploHiC then randomly assigns 'P' or 'M' to the haplotype-unknown end ('chrF, posF') with possibility defined as:

$$possibility(paternal) = \frac{\alpha}{\alpha + \delta} \tag{1}$$

$$possibility(maternal) = \frac{\delta}{\alpha + \delta} \tag{2}$$

Moreover, if the local regions are still unphased after iterative extension, i.e., the sum of contacts counts of haplotype combinations is still zero, HaploHiC will assign the haplotype with uniform possibility depending on the mapped chromosome situation. In this example, if 'chrF, posF' and 'chrL, posL' belong to same chromosome, HaploHiC assigns the haplotype of 'chrL, posL' (the haplotype-known end) to 'chrF, posF' (the haplotype-unknown end). However, if 'chrF, posF' and 'chrL, posL' belong to different chromosomes, uniform possibility (0.5) will be applied.

If intra-chromosome mapped (in this example):

$$possibility(paternal) = 1 \tag{3}$$

$$possibility(maternal) = 0 \tag{4}$$

If inter-chromosome mapped:

$$possibility(paternal) = 0.5 \tag{5}$$

$$possibility(maternal) = 0.5 \tag{6}$$

Hi-C pairs from sEnd-P/M categories are marked as 'phased imputed' and 'unphased imputed' corresponding to phased and unphased local regions, respectively.

B.7 Assign haplotype to unphased Hi-C PE-reads (dEnd-U)

Before the operations on unphased Hi-C PE-reads from dEnd-U category, HaploHiC records 'phased imputed' Hi-C pairs from sEnd-P/M categories to expand the contacts dictionary. For dEnd-U Hi-C pairs, calculation of local contacts counts of haplotype combinations is identical to that of sEnd-P/M mentioned above. Note that as both ends are haplotype-unknown, no haplotype combination will be excluded. Based on local regions' contacts count (α ('P-P'), β ('M-M'), γ ('P-M'), and δ ('M-P'), **Figure 4B**), HaploHiC randomly assigns a haplotype combination to dEnd-U Hi-C PE-reads with possibility defined as:

$$possibility(paternal, paternal) = \frac{\alpha}{\alpha + \beta + \gamma + \delta} \tag{7}$$

$$possibility(maternal, maternal) = \frac{\beta}{\alpha + \beta + \gamma + \delta} \tag{8}$$

$$possibility(paternal, maternal) = \frac{\gamma}{\alpha + \beta + \gamma + \delta} \tag{9}$$

$$possibility(maternal, paternal) = \frac{\delta}{\alpha + \beta + \gamma + \delta} \tag{10}$$

Similar to sEnd-P/M, if the local regions are still unphased after iterative extension, HaploHiC will assign a haplotype with uniform possibility depending on the mapped chromosome status.

If intra-chromosome mapped:

$$possibility(paternal, paternal) = 0.5 \tag{11}$$

$$possibility(maternal, maternal) = 0.5 \tag{12}$$

$$possibility(paternal, maternal) = 0 \tag{13}$$

$$possibility(maternal, paternal) = 0 \tag{14}$$

If inter-chromosome mapped:

$$possibility(paternal, paternal) = 0.25 \tag{15}$$

$$possibility(maternal, maternal) = 0.25 \tag{16}$$

$$possibility(paternal, maternal) = 0.25 \tag{17}$$

$$possibility(maternal, paternal) = 0.25 \tag{18}$$

Hi-C pairs from dEnd-U category are also marked as ‘phased imputed’ and ‘unphased imputed’ corresponding to phased and unphased local regions, respectively.

B.8 Allele-specific integrated results of Hi-C PE-reads

After processing the unphased Hi-C PE-reads of sEnd-P/M and dEnd-U categories, HaploHiC successfully assigns haplotype to all valid Hi-C pairs. Finally, HaploHiC integrates Hi-C pairs of each haplotype combination: intra-paternal ('P-P'), intra-maternal ('M-M'), and inter-haplotype ('P-M' and 'M-P'). Note that 'P-M' and 'M-P' are recorded with different tags in one file.

- Intra-paternal includes dEnd-P category, and imputed ('P-P') Hi-C pairs from sEnd-P/M and dEnd-U categories
- Intra-maternal includes dEnd-M category, and imputed ('M-M') Hi-C pairs from sEnd-P/M and dEnd-U categories
- Inter-haplotype includes dEnd-I and sEnd-I categories, and imputed ('P-M' and 'M-P') Hi-C pairs from sEnd-P/M and dEnd-U categories

All integrated files are in BAM format, which keeps the original alignments of Hi-C pairs. HaploHiC adds several tags in SAM optional fields to denote processing details. A report is generated recording statistics of each category of all Hi-C pairs.

We calculated the phased rate in each sample (**Supplementary Table 10**). The phased rate is the percentage of phased Hi-C pairs, including dEnd-P/M/I, sEnd-I, and phased Hi-C pairs with imputed haplotype from sEnd-P/M and dEnd-U categories. Phased rates in sEnd-P/M and dEnd-U categories are calculated separately.

Additionally, we introduced 'inter-chr imbalance' to evaluate the difference between inter-haplotype and intra-haplotype assignment of inter-chromosome Hi-C pairs. Theoretically, for inter-chromosome contacts, there is no reason to assume any difference between the intra-haplotype and inter-haplotype. The 'inter-chr imbalance' is defined as the difference of intra-haplotype and inter-haplotype inter-chromosome contacts divided by their larger one. Our data shows very low 'inter-chr imbalance' (0.01%-0.04%, **Supplementary Table 10**), which supports the accuracy of HaploHiC phasing.

B.9 Simulation of allele-specific contacts

To evaluate our local contacts based algorithm, we simulated Hi-C sequencing data from haplotype specific contacts between gene pairs of ten categories:
Five categories for imitation of intra-haplotype autosome contacts:

- Cross-Chrom: inter-chromosome translocations
- Long-Distance: intra-chromosome, but on different arms
- Long-Distance: intra-chromosome, on same arm, gene distance is $>$ 10 Mb
- TAD level: intra-chromosome, on same arm, gene distance is [1 Mb, 2 Mb]
- LOOP level: intra-chromosome, on same arm, gene distance is $<$ 700 kb

Two categories for imitation of inter-haplotype contacts:

- InterHap/interChr: inter-chromosome translocations
- InterHap/intraChr: intra-chromosome, gene distance is $>$ 10 Mb

Three categories for imitation of chrX specific activation (intra-haplotype):

- chrX/LongDistance: gene distance is $>$ 10 Mb
- chrX/TAD: gene distance is [1 Mb, 2 Mb]
- chrX/LOOP: gene distance falls into is $<$ 700 kb

In total, 67 gene pairs were randomly selected from COSMIC cancer gene census database (*65*) to construct pairwise allele-specific contacts of maternal and paternal genomes respectively (**Supplementary Table 11**). The maternal and paternal genomes are downloaded from AlleleSeq database (version: Jan-7-2017) (*7*). The simulations on the LOOP level are CN-based (copy number), and the others are SV-based (structure variation, **Supplementary Figure 5**).

To imitate SV-based chromatin contacts for one pair of genes, e.g. *BCR* and *ABL1*, breakpoints are randomly picked from the two genes' genomic regions respectively. At the breakpoints, reciprocal translocations are formed via concatenating extended flanking 2Mb genomic sequences. As the gene-pair contact is heterozygous, sequences are extracted from paternal genome FASTA file to form an allele-specific SV, and sequences from maternal genome are extracted and kept unchanged (**Supplementary Figure 5A**). To simulate CN-based chromatin contacts on the LOOP level, we assign different copy numbers to genomic regions harboring two neighbor genes on the two haplotypes (**Supplementary Figure 5B**). For one genomic region containing two nearby genes (gene *C* and *D*), to make the paternal genome have more contacts between these two genes, we use two copies of this region from the paternal genome, while only keeping one copy from the maternal genome. We applied simu3C (*66*) to simulate the Hi-C sequencing reads of the SV-based and CN-based allele-specific gene-pair contacts.

HaploHiC successfully recovers the allele-specific contacts in simulation data with 97.66% accuracy (**Supplementary Table 11**). First, all SV-based intra-haplotype allele-specific contacts are precisely reported by HaploHiC. For example, contacts of gene pair '*ETV6*, *NTRK3*' are all 'P-P', and contacts of gene pair '*KCNJ5*, *LMO2*' are all 'M-M'. Second, for CN-based intra-haplotype allele-specific contacts, the advantage haplotype is successfully reported by HaploHiC. For example, gene pair '*FLT3*, *LNX2*' has 'P-P' contacts more than three times of the 'M-M' contacts, and gene pair '*RUNX1*, *SMIM11*' has 'M-M' contacts more than two times of the 'P-P' contacts. These ratios are close to what we found in simulation. Third, all inter-haplotype gene pairs are successfully identified with correct the haplotype combination. For example, all contacts of the gene pair '*WT1*, *ZMYM2*' link paternal *WT1* and maternal *ZMYM2*. The gene pair '*MAP2K1*, *NUP214*' has seven 'M-P' contacts (false positives), which is only 2.6% of all its contacts (273). Fourth, all the intra-haplotype chrX allele-specific activation cases are reported correctly. Two gene pairs have some bias to inactivated haplotype ('*LAS1L*, *ZC4H2*' and '*GPC3*, *HS6ST2*'). Note that even shifted, they still show enrichment on correct haplotype. Gene pair '*LAS1L*, *ZC4H2*' has larger shifting (33.3%) to 'M-M' con-

tact, because the neighbor gene (*MSN*, gene distance is about 75 kb) of *LAS1L* forms gene pair '*MSN*, *STAG2*' which is simulated to have only 'M-M' contacts. '*LAS1L*, *ZC4H2*' is influenced by '*MSN*, *STAG2*' as we merged all simulated Hi-C sequencing data of 67 gene pairs together as one sample in HaploHiC evaluation. Finally, all Hi-C pairs with mistakenly assigned haplotype combinations are gathered as false positives (count is 1,326, total count of valid Hi-C pairs is 55,013).

C Whole Chromosome Probe Generation and 3D FISH

Whole chromosome paint probes were generated in-house using PCR labeling techniques as described at https://ccr.cancer.gov/Genetics-Branch/thomas-ried. Chromosome 7 was labeled with Orange dUTP (Abbott Laboratories, Abbott Park, IL), Chromosome 8 was labeled with Dy505 (Dyomics,Jena, Germany) and Chromosome 11 was labeled with Biotin-16-dUTP (Roche Applied Science, Indianapolis, IL). Cells were grown on slides and fixed with 4% paraformaldehyde for 10 minutes. Cells were then washed with 0.05% Triton X100 for five minutes followed by permeabilization steps which included incubation with 0.5% Triton X100 for 20 minutes, followed by subsequent repeated (4x) freeze thaw in liquid nitro- gen/glycerol. The slides were then incubated in 20% glycerol for at least one hour before being frozen in 1xPBS at -20°C until hybridization was performed. Prior to hybridization, cells were washed in 0.05% Triton X100 followed by incubation in 0.1N HCl for 10 minutes. Cells were then washed in 2XSSC followed by incubation in 50% formamide/2XSSC for at least one hour before hybridization. Cells and probes were co-denatured at 72°C for five minutes followed by a 48 hours hybridization at 37°C. After incubation at 37°C, detection commences with posthybridization washes followed by incubation in blocking (3% BSA, 4xSSC, 0.1% Tween20) for 30 minutes at 37°C. The biotinylated probes were detected with the fluorochrome Cy5 conjugated to Streptavidin (Rockland, Gilbertsville, PA). The slides were then washed with 2XSSC before being counterstained with Prolong Gold antifade reagent with DAPI (Promega Madison, WI).

D Spectral clustering: The Laplacian, Fiedler value, and Fiedler vector

The Laplacian, Fiedler value, and Fiedler vector can be summarized as follows. Consider an adjacency matrix $\mathbf{A}$, where $(\mathbf{A})_{i,j} = w(n_i, n_j)$, and weight function, w, satisfying $w(n_i, n_j) = w(n_j, n_i)$ (symmetrical) and $w(n_i, n_j) \geq 0$ (nonnegative). The Laplacian, $\mathbf{L}$, of $\mathbf{A}$ is defined to be $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D} = diag(d_1, \cdots, d_k)$ and $d_i = \Sigma_{j=1}^{k} a_{ij}$. The normalized Laplacian is the matrix $\overline{\mathbf{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$. The second smallest eigenvalue of $\mathbf{L}$ (or $\overline{\mathbf{L}}$) is called the Fiedler value, and the corresponding eigenvector is called the Fiedler vector (*67*). The Fiedler value is also known as the algebraic connectivity of a graph. The magnitude of the Fiedler value increases as the number of edges in the graph increases, and as the graph becomes more structurally ordered. The Fiedler vector partitions the genome into two parts that reflect underlying topology, as given by edge weights inferred from Hi-C data. The Fiedler vector plays a role similar to the eigenvector associated with the largest eigenvalue (principal component 1) of the correlation matrix of the normalized Hi-C matrix (*32*), but it is directly related to properties of the associated graph (*67*). The Fiedler vector can be calculated recursively to identify smaller partitions in Hi-C data. These smaller partitions correspond to topologically associated domains (TADs) (*35*). Description of the Laplacian, Fiedler value, and Fiedler vector adapted from (*19*).

E Structure-Function Visualization and the 4D Nucleome

The goal of the maternal and paternal 4DN is to understand the relationship of allele-specific genomic structure and gene function through time. Gene expression (RNA-seq) directly offers us a scalar value to represent the function of an individual gene, or for a genomic locus (total gene expression for all genes within the locus). In order to find a compatible scalar value which represents structure, we look to Hi-C contacts. Hi-C provides insight into the structural organization of the genome by finding distant regions of the genome (in terms of genomic sequence) that are close to one another in 3D space. The 3D structure of the genome can be viewed through the perspective of a 'genomic network'. In such a network, genomic regions are considered nodes and the contacts between genomic regions are the edges.

A concept that is well known in network theory is centrality. Network centrality measurements, or features, encompass a wide range of network properties (*46*). The common goal of these features is to assign quantitative measurements to the structure of the network. One particularly important feature is the eigenvector centrality, which is the eigenvector associated with the largest eigenvalue of the adjacency matrix which defines the network. Eigenvector centrality assigns values to each node in a network corresponding to that node's influence on the network. Other centrality features that have been shown to have biological relevance include degree, betweenness, and closeness centrality (*48*). We calculated these four centrality features from the Hi-C data at 1 Mb resolution and concatenated them with RNA-seq to form a new structure-function (S-F) matrix which represents both structure and function (rows correspond to genomic loci, columns correspond to centrality features and RNA-seq) (*39, 48*). We derived the S-F matrix both genome-wide and for individual chromosomes. We then normalize the S-F matrix for each setting (maternal and paternal in G1, S, and G2) and concatenate all settings. Next, we apply t-SNE to the combined S-F matrix (containing all settings) and reduce it to two dimensions (*68*). Then we can visualize the two dimensional projection, and observe how the maternal and paternal genomes compare across cell cycle phases (**Figure 9**).

The structure-function matrix integrates genomic structure and gene expression into a common subspace, but it does not directly inform us of the relationship between structure and function. To visualize this allele-specific relationship, we use eigenvector centrality and mature RNA (RNA-seq) for each allele at each phase of the cell cycle. In other words, eigenvector centrality and RNA-seq serve as structure (x-axis) and function (y-axis) coordinates across the cell cycle, respectively (**Figure 6**). The cell cycle and B cell receptor signaling genes displayed in **Figure 6** are contained in a larger sub-network of genes based on their KEGG pathway, whose Hi-C contacts (1 Mb resolution) define the edges of their respective sub-networks (from which eigenvector centrality is calculated). Similarly, the MAE genes displayed are contained within a sub-network of all MAE genes identified in our ABE analysis. The allele-specific genes displayed, which serve as a control, are contained within a gene sub-network which was randomly

selected from all allele-specific genes. The number of allele-specific genes selected for the control is the mean of the other three sets' sizes. The three points (G1, S, and G2) in the 4DN phase plane for each allele are fit with a minimum volume ellipse to capture the 4DN variance of the allele (*69*). Allelic divergence for each gene is calculated as the average distance between the maternal and paternal alleles' S-F coordinates for G1, S, and G2, after normalization of coordinates. The allelic divergence for each group of genes displayed in **Figure 6** is defined as the mean of these genes' individual allelic divergences.

F Structure Alignment and 3D Modeling

Structures computed from different distance matrices could be varied in both scale and orientation. To align different structures and superimpose them into the same coordinates, we used Procrustes analysis (*70*) which applies the optimal transform to the second matrix (including scaling/dilation, rotations, and reflections) to minimize the sum of square errors of the pointwise differences. The first step of Procrustes analysis is to translocate shapes to the origin by subtracting the mean value of all coordinates. The next step is to force shapes into the same scale by dividing each shape by the Frobenius norm. For an $m \times n$ matrix $\mathbf{A}$, the Frobenius norm is defined as $\|\mathbf{A}\|_F = (\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2)^{\frac{1}{2}}$. Finally, we find an optimal rotation matrix that will align one contact matrix $\mathbf{A}$ to matrix $\mathbf{B}$ by using Singular Value Decomposition (SVD) on $\mathbf{M}$ where $\mathbf{M} = \mathbf{A}^\top \mathbf{B}$. Applying SVD to the matrix $\mathbf{M}$ gives $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{V}^\top$ is the rotation matrix of $\mathbf{B}$. Then $\mathbf{B}\mathbf{V}^\top$ is the rotated shape. To visualize the final structure from the inferred three dimensional embedding, we smooth the curve by interpolating three dimensional scatter data using the radial basis function (RBF) kernel. The structure was then visualized using the Mayavi package in Python (*71*).

G Statistical Significance via Permutation Test

A permutation test builds the shape of null hypothesis (namely, the random background distribution) by resampling the observed data. In our analysis, this sampling procedure was repeated

10,000 times, and a rank-based p-value is then calculated for the right-tailed event for testing likelihood of structure changes around ABE genes and the left-tailed event for decreased ABE in an essential gene set (*31*). The background distribution for ABE in essential genes was generated by calculating the proportion of ABE genes in a randomly selected set of 698 allele-specific genes (same number of genes as the allele-specific essential gene set). The proportion of ABE genes in the essential gene set compared to this background distribution yields the p-value (**Supplementary Figure 3**). The background distributions for genome structure changes were generated by calculating the average number of significant changes in the Hi-C contacts surrounding random allele-specific genes. The probability of the right-tailed event of our observation of significant structural changes around ABE, CBE, and combined set of ABE and CBE genes under their respective background distributions yields the p-values (**Supplementary Figure 7**).

Supplementary Figures

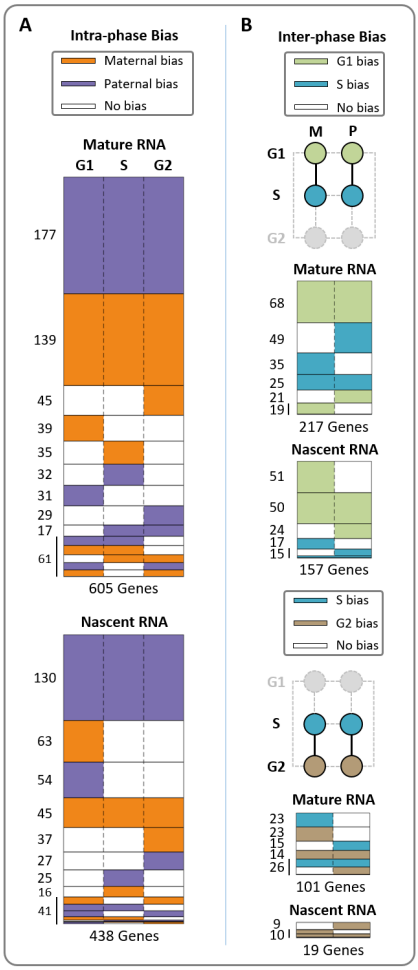


Fig. 1: Expression biases identified between alleles and cell cycle phases. **(A)** Differential expression between the maternal and paternal alleles for each cell cycle phase. Differentially expressed genes with a bias towards the maternal (paternal) allele in a particular cell cycle phase are shaded orange (purple). Each row represents a group of genes with a specific expression bias profile. Analogous to **Figure 1**. **(B)** Differential expression between cell cycle phases within each allele. Top shows differential expression between G1 and S phases while bottom shows differential expression between S and G2. Genes are shaded in the color of the cell cycle phase in which their expression is significantly higher. G1, S, and G2 are shaded green, blue, brown respectively. In both (A) and (B), white represents no expression bias.

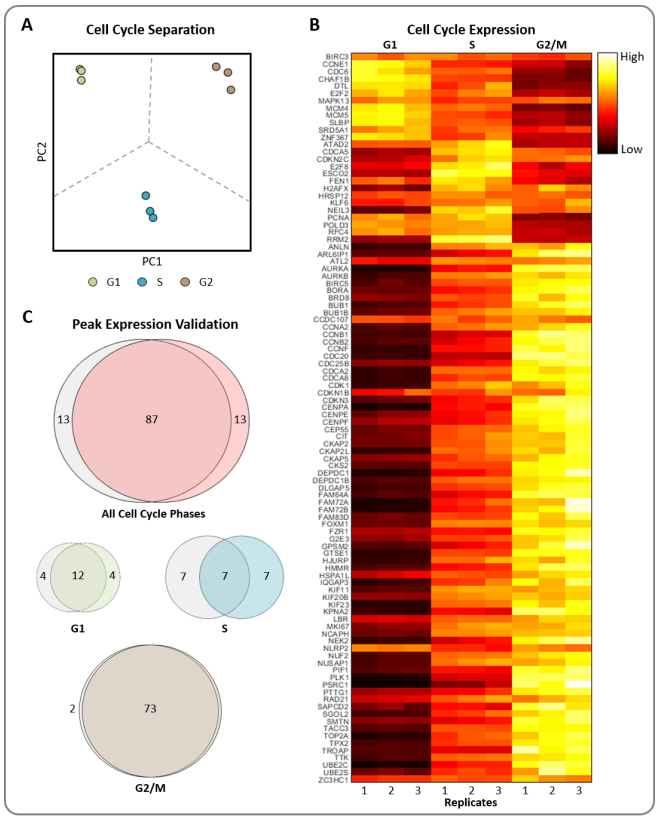


Fig. 2: Cell cycle separation of RNA-seq data. **(A)** Principal component analysis (PCA) of three replicates over three time points. G1, S, and G2 are easily separated and the replicates of each cell cycle phase cluster after dimension reduction. **(B)** RNA-seq expression heatmap of the top 100 periodically expressed genes in our data from Cyclebase are row normalized and clustered by peak expression time (*72*). **(C)** Venn diagram comparison for the peak expression time of the same 100 periodically expressed genes. Red, green, blue, and brown circles represent the peak expression times for all cell cycle phases, G1, S, and G2 within our data respectively, and overlap with gray indicates agreement with Cyclebase peak expression times. Over all phases, 87% of these genes match in their peak expression cell cycle phase.

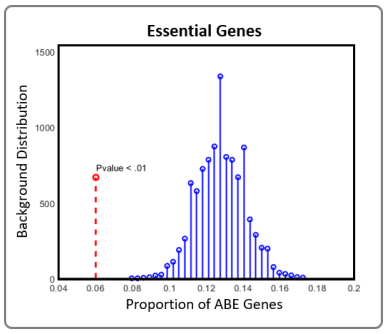


Fig. 3: Statistical significance of decreased ABE in essential genes. A permutation test was performed to determine whether essential genes were less likely to have ABE versus randomly sampled allele-specific genes (**Supplementary Notes G**).

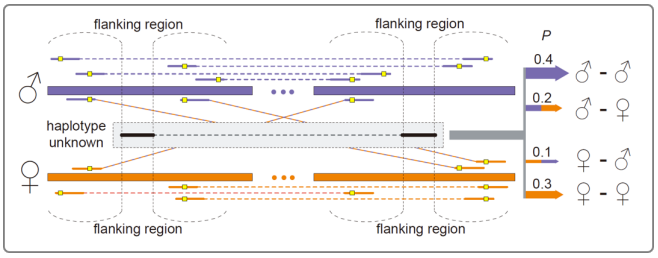


Fig. 4: Expanded diagram of HaploHiC phasing in Figure 4. Haplotype-unknown Hi-C pair is distributed to certain haplotype combination with probabilities depending on local phased contacts in flanking regions. Homologous genomic regions are linked by phased Hi-C pairs (purple for paternal, orange for maternal). Phased heterozygous loci are denoted by yellow squares.

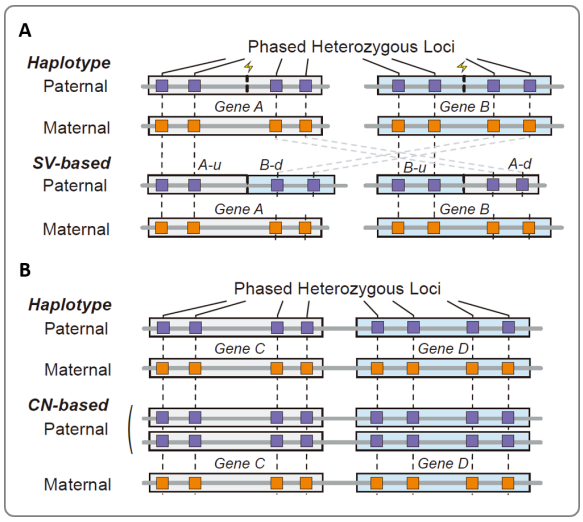


Fig. 5: Simulation of SV-based and CN-based contacts. **(A)** Gene pairs are reciprocally partially concatenated on the paternal haplotype to form a chimeric sequence. The maternal haplotype carries the wild-type. Upstream and downstream genomic regions are denoted by 'u' and 'd', respectively. **(B)** The copy number of gene pairs' genomic region on the paternal haplotype is increased, and the maternal haplotype is kept at one copy.

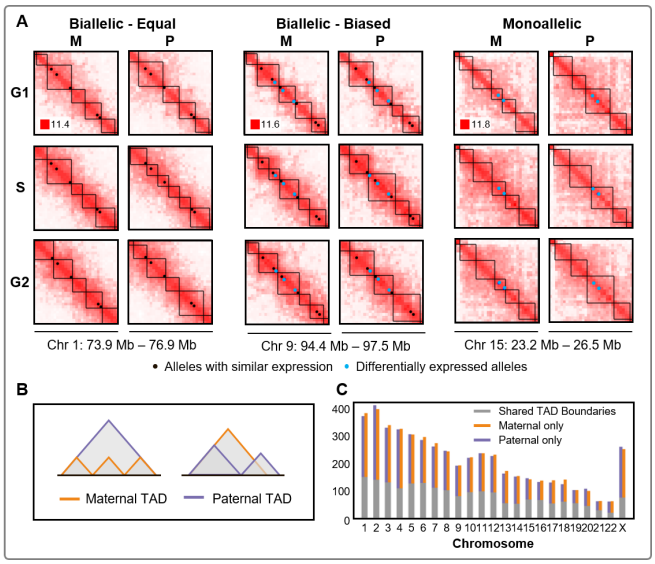


Fig. 6: Allele-specific topologically associated domains (TADs). (**A**) Subsection of Chromosomes 1 (left), 9 (middle), and 15 (right) separated into allele- and cell cycle-specific matrices shown in log_2 scale 100 kb resolution, with TADs depicted as solid black lines. Circles in each setting represent allele-specific genes with similar (black) and differential (blue) expression. TAD boundaries are found using the spectral identification method (*35*). **(B)** Schematic representation of allele-specific TADs. A single TAD in one genome can be identified as a multiple TADs in the other genome (left), or TAD boundaries can shift depending on local genome structure (right). **(C)** Bar graph highlighting the number of TAD boundaries in common between maternal and paternal (gray), maternal only TAD boundaries (orange), and paternal only TAD boundaries (purple) for all chromosomes in G1.

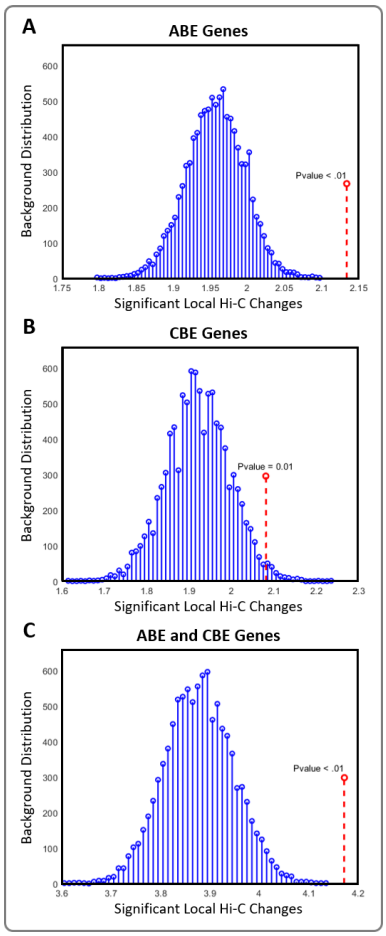


Fig. 7: Statistical significance of changes in chromatin structure around DE genes. Permutation tests were performed for ABE genes, CBE genes, and the combined set of ABE and CBE genes to determine whether significant local structure changes were more likely in DE genes versus randomly sampled allele-specific genes for their respective comparisons (**Supplementary Notes G**).

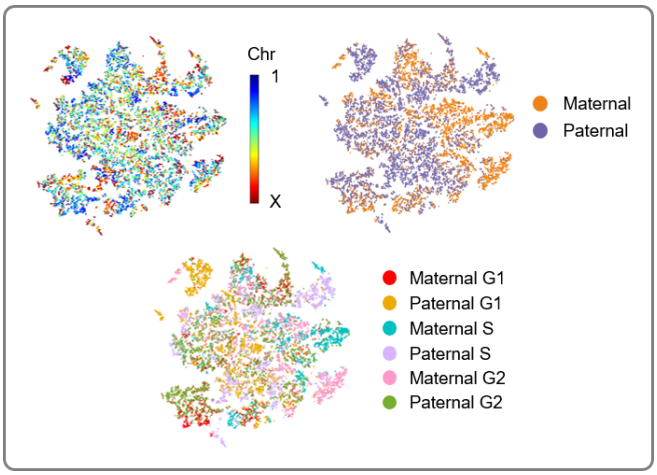


Fig. 9: Genome-wide visualization of the maternal and paternal 4DN. Low dimensional representation (t-SNE) of structure and function for the maternal and paternal genomes across cell cycle phases (1 Mb resolution) (**Supplementary Notes E**) (*68*). The three plots show the same data with different color schemes to highlight the locations of chromosomes (top left), parental origin (top right), and parental origin with cell cycle phase (bottom). Differences in the low dimensional mappings can be seen genome-wide for the maternal and paternal genomes, as well as cell cycle phases.

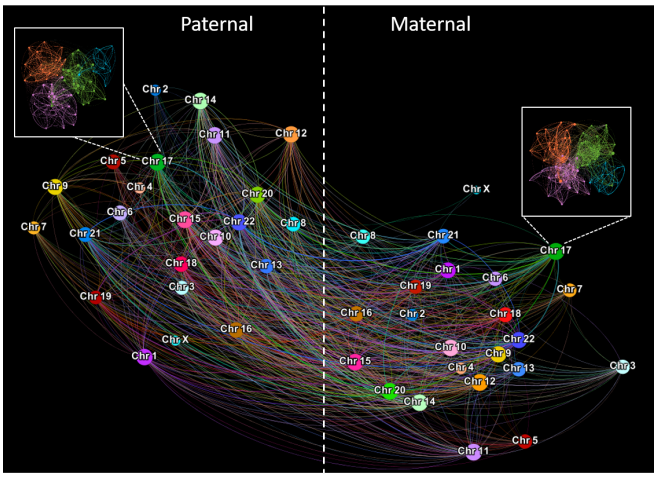


Fig. 8: Maternal and paternal genomes as a network of chromosomes. The maternal and paternal chromosomes are separated to highlight inter-haplotype interactions from Hi-C data. Each chromosome has an internal network of intra-chromosomal chromatin contacts as highlighted in the maternal and paternal Chromosome 17.

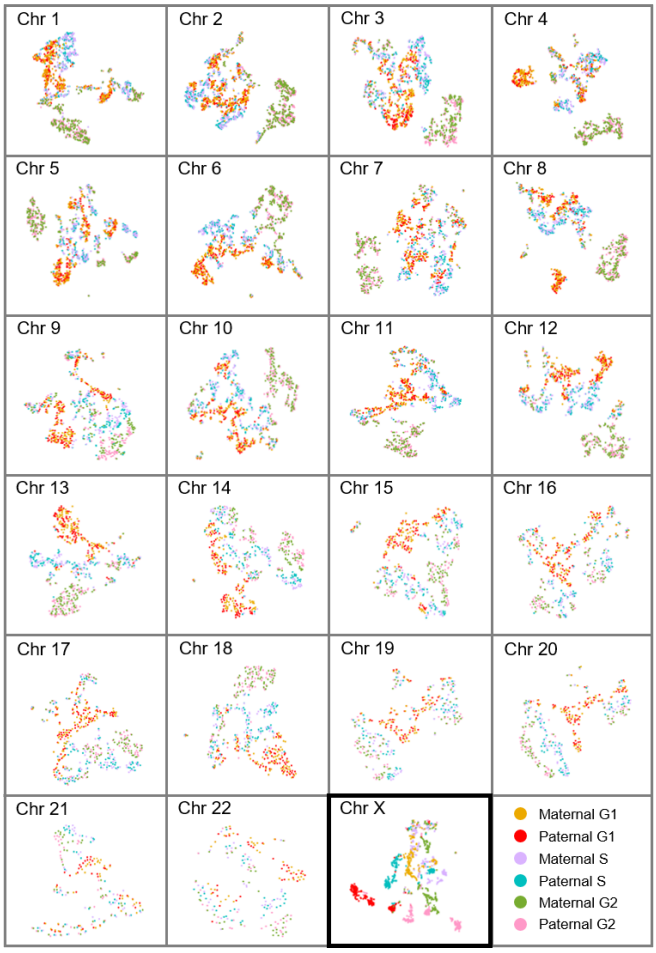


Fig. 10: The same process from **Supplementary Figure 9** was applied to individual chromosomes, with color scheme highlighting parental origin and cell cycle phase. The X Chromosome shows clear separation of the maternal and paternal copies. The autosomes have more subtle differences between their maternal and paternal low dimensional projections.

Supplementary Tables

Supplementary Table 1: Differential expression per chromosome. Number of allele-specific genes, DE genes between alleles (ABE) and cell cycle phases (CBE), and percentage of allele-specific genes that are DE for each chromosome.

Supplementary Table 2: Differential expression between alleles using RNA-seq. ABE was observed for 605 genes between the maternal and paternal alleles in mature RNA.

Supplementary Table 3: Differential expression between alleles using Bru-seq. ABE was observed for 438 genes between the maternal and paternal alleles in nascent RNA.

Supplementary Table 4: Differential expression between cell cycle phases using RNA-seq. CBE was observed for 284 genes between cell cycle phases for the maternal or paternal alleles in mature RNA.

Supplementary Table 5: Differential expression between cell cycle phases using Bru-seq. CBE was observed for 171 genes between cell cycle phases for the maternal or paternal alleles in nascent RNA.

Supplementary Table 6: Coordination of allele-biased TF binding and gene expression. Genes are separated into three groups for Pol II, CTCF, and other TFs. Each gene contains the number of binding instances, the parental bias, and whether or not the bias had agreement with their ABE.

Supplementary Table 7: Allele biased expression of KEGG pathways. All KEGG pathway gene sets with at least five allele-specific genes were examined for ABE. Here, we list the the percentage of allele-specific genes with ABE for each pathway.

Supplementary Table 8: Details of phased germline mutations of GM12878. Mutation counts of diverse allele scenarios on maternal and paternal genomes are listed.

Supplementary Table 9: Parental origin categories of Hi-C PE-reads in HaploHiC. Seven Hi-C pairs categories are denoted with judgment conditions of paired-end alignments.

Supplementary Table 10: Details of Hi-C PE-reads in sample GM12878 processed by HaploHiC. Hi-C pair counts of all categories in three steps are listed.

Supplementary Table 11: Details of ten categories of gene pairs in Hi-C simulation and evaluation of HaploHiC. In total, 67 gene pair details are listed with parental and maternal origin contacts division results from HaploHiC.