Genome architecture and transcription data reveal allelic bias during the cell cycle

Stephen Lindsly[1], Wenlong Jia[2], Haiming Chen[1], Sijia Liu[3], Scott Ronquist[1], Can Chen[4,7], Xingzhao Wen[5], Gilbert 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

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

Highlights

- Structural and functional differences between the maternal and paternal genomes, including similar allelic bias in related subsets of genes.
- Coupling between the dynamics of gene expression and genome architecture for specific alleles illuminates an allele-specific 4D Nucleome.
- Introduction of a novel allele-specific phasing algorithm for genome architecture and a quantitative framework for integration of gene expression and genome architecture.

Abstract

Millions of genetic variants exist between paternal and maternal genomes in human cells, which result in unequal allelic contributions to gene transcription. We sought to understand differences in transcription and genome architecture between the maternal and paternal genomes across the cell cycle. Our approach integrates haplotype-resolved genome-wide data from B-lymphocytes in G1, S, and G2/M, mature and nascent mRNA, genome wide chromosome conformation capture (Hi-C), and protein binding (ChIP-seq). Further, our novel method, HaploHiC, enables phasing of reads of unknown origin. Our data indicate significant allelic bias in cell identity genes, and notably absent bias in universally essential cell cycle and circadian rhythm genes. We offer a model in which coordinated biallelic expression reflects prioritized preservation of essential gene sets.

Introduction

Maternal and paternal chromosomes typically each encode an allele of the same gene. Millions of single nucleotide variants (SNVs), insertions, and deletions (InDels) enable distinction between the maternal and paternal genome, and DNA sequence variations can influence 3D conformation of the genome (*1*). Studying the maternal and paternal alleles of a gene may enhance our understanding of genome organization and mechanisms of bias in allele expression. Allelic differences have been observed through gene expression, 3D chromatin organization, and protein binding (*2–9*), and it has been shown that even subtle differences in genome structure can lead to differential gene expression and regulation (*10*). Approximately 10-25% of all genes have allelic biased expression across many cell types (*11–13*). One well studied example of allele-specific expression and architecture is X chromosome inactivation (XCI) in females, in which one of the two X chromosomes is tightly packed into two 'superdomains' with silencing of gene expression (*4, 14*). Another example of known allele-specific expression is imprinting, in which one allele's expression is suppressed (*15*). It has also been shown for single cells that monoallelic transcripts could make up over 75% of total transcripts (*16*).

The allele-specific nature of genome architecture and gene expression have been explored individually, but it remains poorly understood how differences in chromosome architecture between alleles contribute to differences in gene expression. To address this problem, we look to transcription factors and other regulatory proteins as key components that connect differences in genome conformation with differential expression of the maternal and paternal genomes. By integrating RNA sequencing (RNA-seq), bromouridine sequencing (Bru-seq), and genome-wide chromosome conformation capture (Hi-C), along with publicly available protein binding data (ChIP-seq), we contribute insight into how maternal and paternal genomes contribute to cell phenotype.

The redundancy inherent in diploid genomes makes them robust to mutations and genetic loss-of-function events. In cancers, this redundancy is often missing, and referred to as loss of heterozygosity. For example, *BRCA1* and *BRCA2* exhibit loss of heterozygosity (*17*), and

mutations in these genes lead to a significant increase in the likelihood of developing breast cancer (*18*). Additionally, monoallelically expressed genes, like X-inactivated and imprinted genes (*15, 19–21*), do not have a 'backup' gene to compensate for potential deleterious mutations, and it has been found that mutations in monoallelicly expressed genes are responsible for disease phenotypes (*22*). These findings lend credence to the pursuit of allele-specific genome analysis.

The 4D Nucleome (4DN) describes the relationship between genome architecture, gene expression, and cell phenotype through time (*23*). Chen *et al.* (*24*) introduced a mathematical framework for the 4DN, which can be applied in an allele-specific manner to a fully haplotyped cell line (e.g. NA12878). Despite extensive study of genome replication and compaction in cell division, few studies focus on differences between maternal and paternal genomes during these processes. For example, the two alleles may have different dynamics through the cell cycle that influence proliferation, or an allele-specific 4DN.

In this paper, we integrate allele-specific RNA-seq, Bru-seq, Hi-C (**Figure 1**), and ChIP-seq data to provide a more comprehensive understanding of the maternal and paternal contributions to phenotype, and the relationship between these properties. We analyze differential expression for allele-biases and allele-specific cell cycle-biases, chromatin compartment switching (also known as A/B compartments), and Topologically Associated Domains (TADs). In addition, we analyze our Hi-C data using our recently developed methods that help discern differences in genome architecture between the two genomes and through the cell cycle. Our methods for analyses reveal notable patterns of allelic bias across the cell cycle, including subsets of genes involved in related cellular processes. This work supports the idea that separate analysis of the two parental genomes is imperative to future genomic studies.

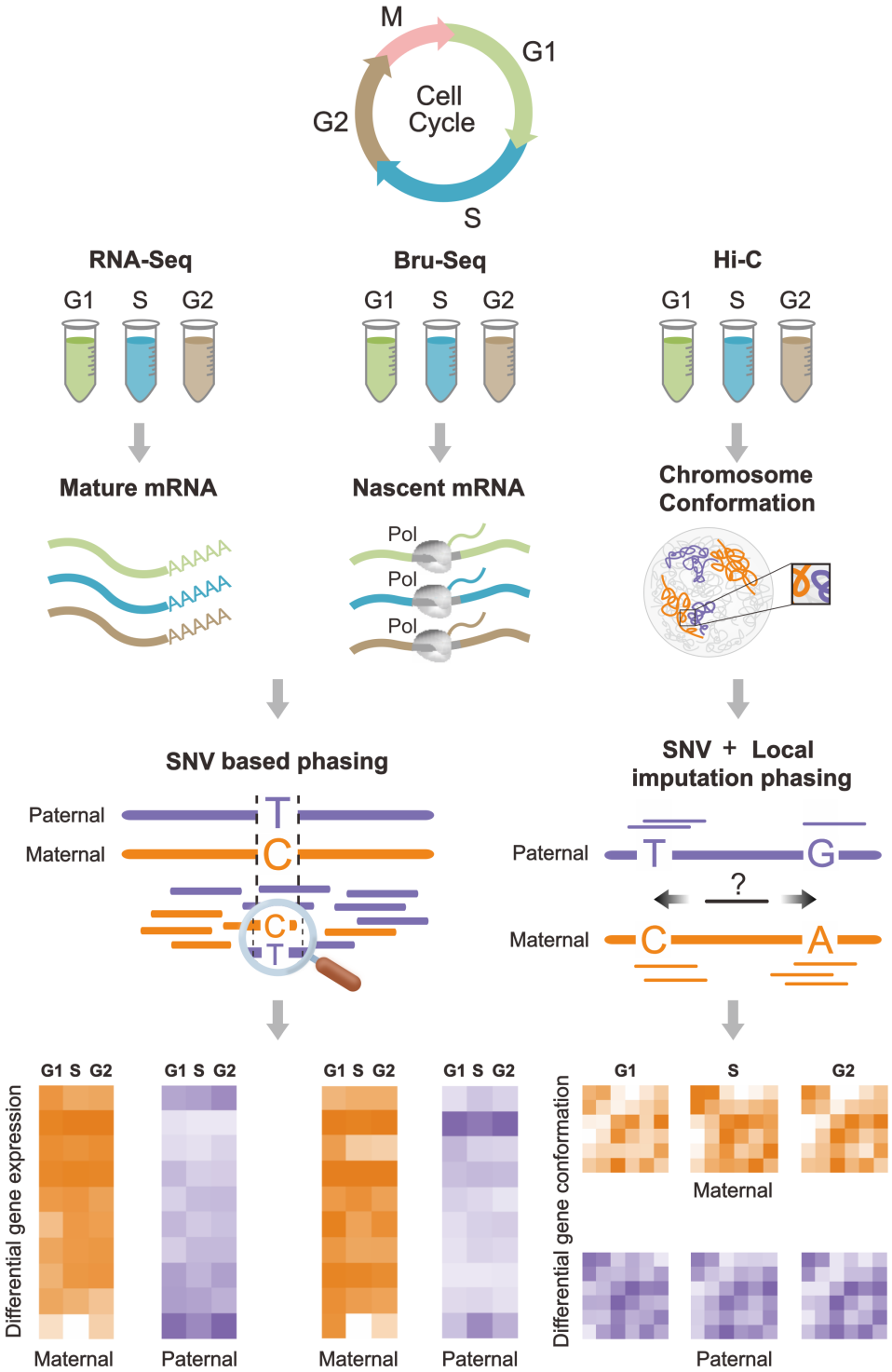

Fig. 1: Experimental and allelic separation workflow. Cell cycle sorted cells were extracted for RNA-seq, Bru-seq, and Hi-C (left to right, respectively). RNA-seq and Bru-seq data were allelicly phased via SNVs/InDels (left). SNV/InDel based imputation and allelic phasing done using HaploHiC (right).

Results

Chromosome level differences of the maternal and paternal genomes

In order to establish whether there are allele-specific differences in both chromosome structure and gene expression, we analyzed parentally phased whole-chromosome Hi-C matrices and RNA-seq data at a 1 Mb scale (phasing discussed below, in **Supplementary Notes 2.1.7**, and **2.4**). We performed a simple subtraction of the phased Hi-C matrices and found the Frobenius norm of the difference matrices, adjusted for chromosome size. Similarly, we subtracted the phased RNA-seq vectors and found the Frobenius norm of each difference vector, adjusted for chromosome size. If the two parents' genomes were equivalent, the resulting chromosome-level Frobenius norms would approach zero. As seen in **Figure 2**, we found that all chromosomes have allelic differences in both structure (blue) and function (red). Unsurprisingly, chromosome X had the largest structural difference, followed by chromosomes 9, 21, and 14. The X chromosome also had the most extreme functional difference, followed by chromosomes 13, 7, and 9.

We standardized the Frobenius norms of all chromosomes by the norm of the X chromosome, as its value was much larger than any of the autosomes. A threshold was assigned at the median Frobenius norm for Hi-C and RNA-seq (**Figure 2** green dashed lines). Interestingly, the majority of chromosomes with larger (smaller) Frobenius norms than the median for Hi-C are also larger (smaller) than the median in RNA-seq. Although there are some exceptions, the majority of large allele-specific structural differences correspond to large functional differences throughout the genome.

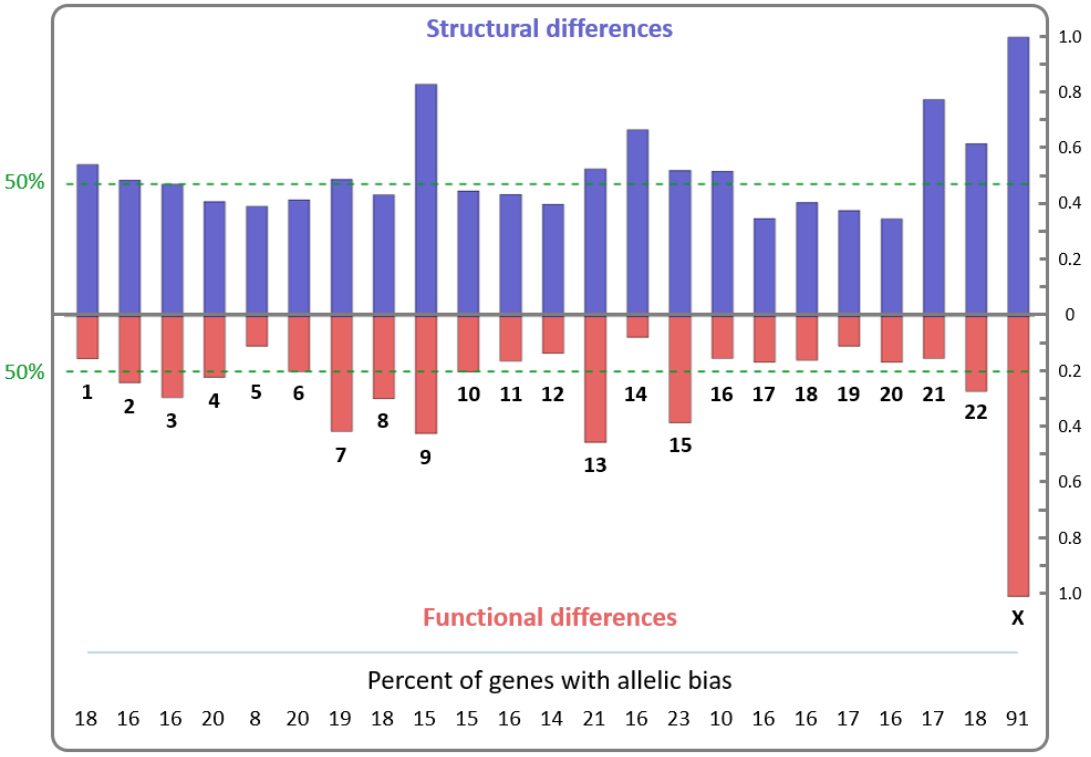

Fig. 2: Chromosome level structural and functional parental differences. Structural differences (blue) represent the aggregate changes between maternal and paternal Hi-C over all 1 Mb loci for each chromosome, adjusted for chromosome size in G1 phase. Functional differences (red) represent the aggregate changes between maternal and paternal RNA-seq over all 1 Mb loci for each chromosome, adjusted for chromosome size in G1 phase. 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. Percent allelic bias was determined by the number of differentially expressed alleles divided by the number of allele-specific genes in each chromosome (additional information in **Supplementary Table 1**).

Allele-specific differential expression dynamics

After confirming the existence of allelic differences on a chromosome level, we investigated all allele-specific gene expression through RNA-seq and Bru-seq. We evaluated the allele-specific genes for differential expression in seven comparisons between the six settings: maternal and paternal in G1, S, and G2 (G2/M referred to as G2 for brevity). These comparisons consist of maternal vs paternal in each of the cell cycle phases, as well as G1 vs S and S vs G2 for the

maternal and paternal genomes, respectively. We then used K-means clustering to sort genes by their patterns of expression. A gene was considered to be differentially expressed if its two alleles' expression levels have an adjusted p-value (P_{adj}) < 0.05 from DESeq (*25*), and a fold change (FC) > 2.

First, we identified genes with allele-biased expression (ABE) and allele-specific cell cycle-biased expression (CBE) from our RNA-seq data. From the 23,277 genes interrogated, we identified 6,795 genes that contained informative heterozygous single nucleotide variations (SNVs), small insertions (In), and deletions (Dels). There were 5,003 genes with read counts ≥ 5 covering SNVs/InDels, which we found to be the minimum number of counts needed to reliably determine allele-specificity. We performed differential expression analysis (*25*) for the seven comparisons to identify which of the 5,003 genes had ABE or CBE. This model identified 778 genes (FDR < 0.05) that had > 0.1 reads per kilobase million (RPKM), and showed significant ABE in both $P_{adj} < 0.05$ (*26*) and fold change > 2. Out of the 778 differentially expressed genes, we identified 605 ABE genes (**Supplementary Table 2**) that were differentially expressed between alleles and 284 CBE genes (**Supplementary Table 4**) that were differentially expressed when comparing cell cycle phases of each allele individually (111 genes in common).

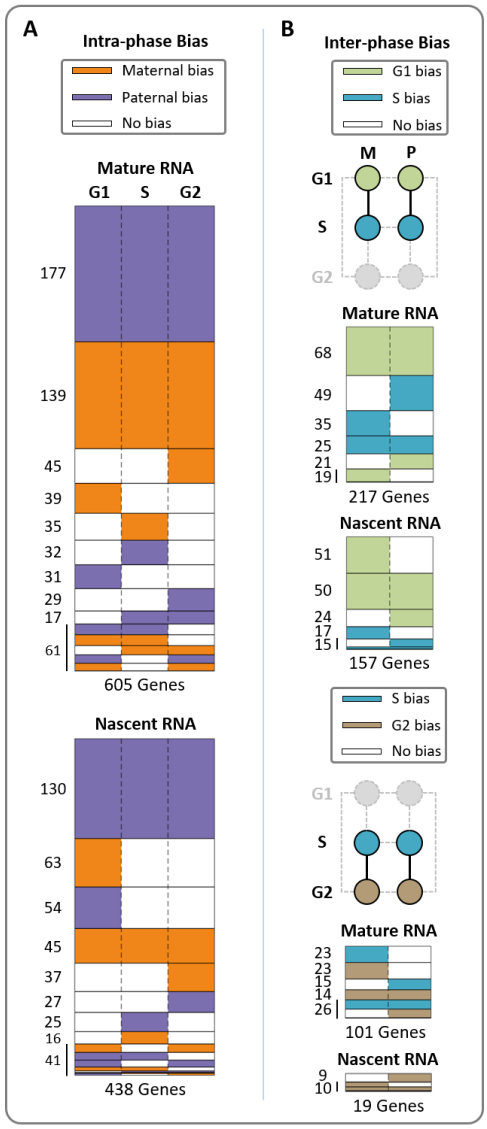

Fig. 3: Expression biases identified between alleles and cell cycle phases. **(A)** Differential expression between the maternal and paternal alleles for each cell cycle phase. Differentially expressed genes with a bias towards the maternal (paternal) allele in a particular cell cycle phase are shaded orange (purple). Each row represents a group of genes with a specific expression bias profile. **(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.

To determine ABE and CBE genes in our Bru-seq data, both exons and introns containing informative SNVs/InDels were used. We identified 266,899 SNVs/InDels from the Bru-seq data, compared to only 65,676 SNVs/InDels from the RNA-seq data. However, in the Bru-seq data, many SNVs/InDels had read depths that were too low (< 5) to be included in the analysis. Using a read depth criterion ≥ 5, similar numbers of reliable SNVs/InDels were found in the RNA-seq and Bru-seq data (19,394 and 19,998 respectively). Using this set of reliable SNVs/InDels in our Bru-seq data, we were able to find a total of 5,832 genes that were allele-specific. From these allele-specific genes, DESeq identified 556 genes (FDR < 0.05) that had RPKM values > 0.1 and showed significant ABE or CBE. Out of the 556 genes, 438 (**Supplementary Table 3**) were differentially expressed between alleles and 171 (**Supplementary Table 5**) were differentially expressed between cell cycle phases of each parent (53 genes in common).

We separated the differentially expressed genes into their respective chromosomes to observe their distribution throughout the genome. From RNA-seq (Bru-seq), we found that most autosomes had between 15-20% (5-10%) of their allele-specific genes differentially expressed, either between the maternal and paternal alleles or between cell cycle phases (**Figure 2**). As expected, the X chromosome had a particularly high percentage of differentially expressed genes at 90.5% (89.5%). We found that chromosome 15 had the highest percent of differentially expressed genes among autosomes at 22.7% in RNA-seq while chromosome 18 had the highest percent in Bru-seq at 16.4% (**Supplementary Table 1**).

Allele-biased expression

In **Figure 3A**, we focus on genes with differential expression between the maternal and paternal alleles for each cell cycle phase. We found that 316 genes have ABE in all cell cycle phases (177 upregulated in paternal, 139 in maternal) from RNA-seq and 175 (130 paternal high and 45 maternal high) from Bru-seq. In addition to these, 281 genes have differential expression in only one or two cell cycle phases from RNA-seq and 264 from Bru-seq.

The largest clusters of genes with ABE have a bias toward one parent throughout the entire

cell cycle (top two clusters in mature RNA, top and fourth cluster of nascent RNA). These clusters include X-inactivated, imprinted, and other monoallelically expressed genes. Examples of this behavior are highlighted in the 'X-Linked' and 'Imprinted' sections of **Figure 4A**, as well as *TUBB6* and *NPNT* in the 'Autosomal Genes' section. Of note, the *XIST* gene, responsible for X chromosome inactivation, is expressed only in the maternal allele. This deactivates the maternal X chromosome without affecting the expression of the paternal alleles. X-inactivation can also be confirmed by the presence of Hi-C superdomains in the maternal X chromosome, and absence of these superdomains in the paternal X chromosome (**Figure 6E**). Many genes that are differentially expressed between alleles only differ in one or two cell cycle phases. This implies that at certain cell cycle phases, the expression of the two alleles is coordinated, but this coordination is lost at other points in the cycle. This can be seen in many of the 'Autosomal Genes' presented in **Figure 4A**.

Among the ABE genes from RNA-seq analysis, we found 117 monoallelicly expressed genes. In addition to the requirements for differential expression, we impose the thresholds of a FC ≥ 10 and for the inactive allele to have < 0.1 RPKM, or FC ≥ 50 for all cell cycle phases. These 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*). These genes and their related diseases offer further support for allele-specific analysis of both structure and function, as their monoallelic expression could not be detected without allele-specificity.

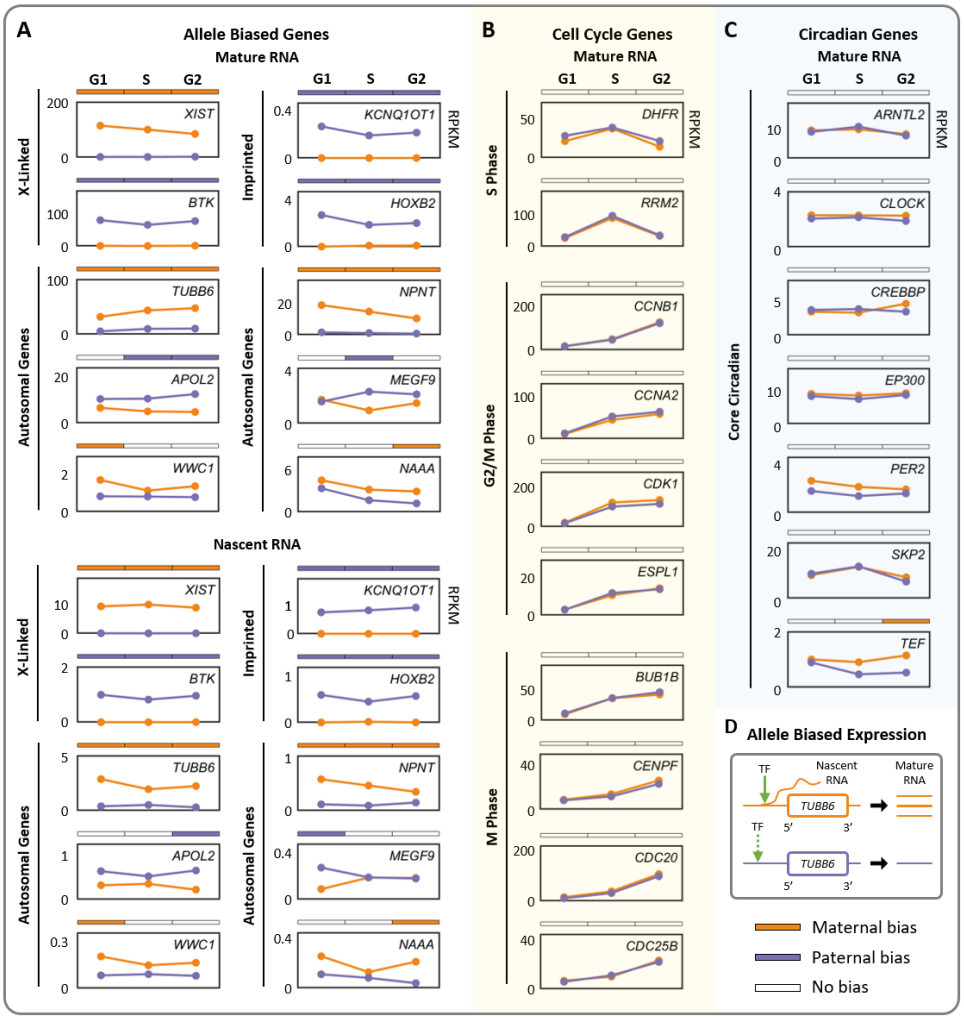

Fig. 4: Allele-specific analysis reveals biased and coordinated gene expression. **(A)** Representative examples of X-linked, imprinted, and other genes with allelic bias. As in **Figure 3A**, genes are shaded orange (purple) when they are biased toward the maternal (paternal) allele. Top shows mature RNA levels of these genes and bottom shows nascent RNA expression. **(B)** 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 cell cycle-biased expression. Notably, these genes happen to have no significant bias between alleles. **(C)** Core circadian genes' mature RNA levels through the cell cycle. Similar to the cell cycle regulatory genes, all but one of these genes show no allelic bias in expression. **(D)** Schematic representation of the gene *TUBB6* which has a large maternal bias in both nascent and mature RNA expression.

Cell cycle-biased expression

In **Figure 3B**, we show the gene expression dynamics through the cell cycle for each of the parent's genomes. The top of **Figure 3B** compares gene expression of each allele between G1 and S, while the bottom compares between S and G2. This figure gives us insight into the differences between maternal and paternal alleles' dynamics through 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 mature RNA expression, 68 genes are downregulated from G1 to S in both parental genomes and 25 are upregulated. In nascent expression, 50 genes are downregulated and 3 are upregulated. In the S to G2 comparison, there are 27 (4) genes in RNA-seq (Bru-seq) that have similar expression dynamics between alleles. From mature RNA, 14 genes are upregulated while 13 are downregulated and in nascent transcription, 4 genes are upregulated while none are downregulated.

From these data, we can clearly see a coordination of expression between many, but certainly not all, alleles through the cell cycle. As most cell cycle related analyses do not discriminate between the maternal and paternal alleles' dynamics, it is not currently understood why alleles have different expression patterns through the cell cycle. We hypothesize that the differences in allele-specific cell cycle dynamics may be related to replication timing (*27, 28*), but further investigation is needed to understand the mechanisms behind these differences.

In **Figure 4A**, we show example genes with various patterns of differential expression between alleles through the cell cycle. Our analysis confirmed allelic biases through X-inactivated and imprinted genes for this cell line, as shown by the genes *XIST*, *BTK*, *KCNQ1OT1*, and *HOXB2*. X-inactivated and imprinted genes are silenced via transcriptional regulation which can be verified by monoallelic nascent RNA expression. Interestingly, the X chromosome that was inactivated in our cells is opposite of what is commonly seen for the NA12878 cell line in literature (*4, 29*) but is consistent between our RNA-seq, Bru-seq, and Hi-C findings.

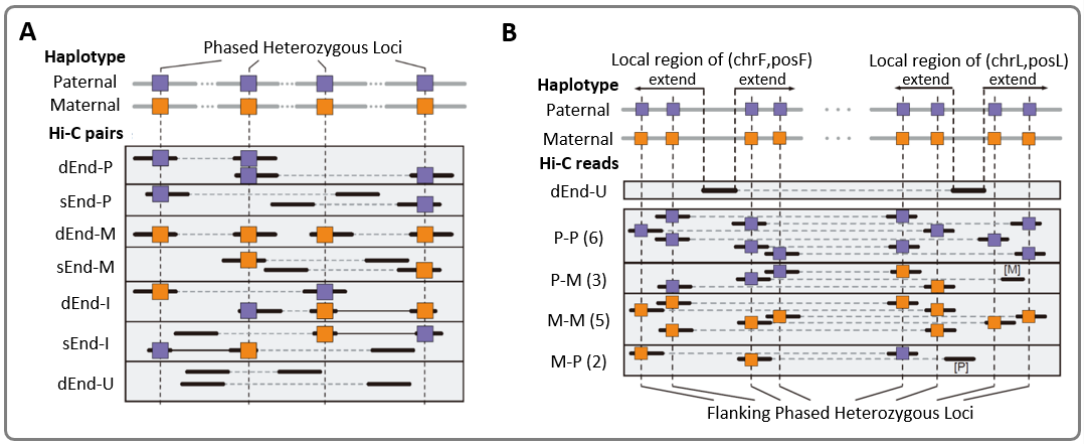

Fig. 5: A novel phasing technique for allele-specific Hi-C data, HaploHiC. **(A)** Paired-end reads are separated into groups based on parental origin determined through SNVs/InDels (see detailed definition in **Supplementary Notes 2.1.4**). Reads are grouped by: (i) reads with one (sEnd-P/M) or both 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). **(B)** An example of a paired-end read (dEND-U) with no SNVs/InDels has its origin imputed using nearby reads (**Supplementary Notes 2.1.7**). 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 2.1.5**). This ratio then determines the likelihood of the haplotype-unknown read's origin.

In addition to known allele-biased genes, like those which are X-inactivated or imprinted, we also found hundreds of genes with allele-biased expression in at least one cell cycle phase. Genes like *TUBB6* and *NPNT* were strongly biased toward the maternal allele in all cell cycle phases. These genes also appeared to be transcriptionally repressed, as the paternal alleles had minimal mature RNA present and heavily biased nascent transcription. This relationship between mature and nascent transcription holds true for many genes, but not all genes with allelic bias in mature transcription have corresponding nascent transcription with a sufficient read depth to be reliable.

Approximately half of all allele-biased genes in mature RNA, and the majority in nascent transcription, had differential expression in one or two cell cycle phases. Four examples of genes with this behavior are shown in the 'Autosomal Genes' section of **Figure 4A**. It is not well understood why alleles would vary in their relative expression through the cell cycle, and

this warrants further investigation in the future.

A survey of genes known to be critical in the regulation of the cell cycle and the circadian rhythm was also performed. As expected, the cell cycle genes were found to be differentially expressed between cell cycle phases. Remarkably, none of the well known cell cycle regulating genes exhibited allelic biases (**Figure 4B**). In the set of core circadian genes, only one gene (*TEF*) had an allelic bias in one cell cycle phase (**Figure 4C**). We hypothesize that the functions of these genes are so crucial for a healthy cell, that both alleles must be active to ensure the cell has at least one working copy, even in the case of mutations or other unforeseen issues. Another possible explanation for the coordination between alleles of cell cycle and core circadian genes is the total level of expression needed for proper function. If these genes need to produce a certain amount of RNA in order to be effective, it would be beneficial to the cell for the alleles to be equally active. If these genes exhibited ABE, the total amount of RNA expression from the two alleles may not satisfy this requirement.

Allelic differences in genome structure through the cell cycle

To examine the 3D structure of chromatin, we use high-throughput genome-wide chromatin conformation capture, or Hi-C. 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 look to separate the maternal and paternal genomes' contributions to this contact matrix to further analyze the allelic differences and their effects on 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 genomic sections that are identical. Due to the cells' ability to correct errors that arise in their DNA, there are a relatively small number of variations that persist within a person. This means that our segregated maternal and paternal contact matrices will be extremely sparse since

we can only definitively tell the difference between the respective genomes at a small subset of genomic loci. In order to combat this problem, we look to reintroduce unphased reads to both the maternal and paternal contact maps.

Here we propose HaploHiC, a novel technique for phasing reads of unknown origin using a local imputation of known reads.Unlike previous haplotyping methods (*30*), 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 5, Supplementary Notes 2.1.4-2.1.7**). For example, we would expect that in an area with many reads mapped to the paternal genome, unphased reads in that area 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 10, Supplementary Notes 2.1.9**). The allelicly separated genomes can be seen in **Figure 6A**, with selected chromosomes highlighted in **Figure 6B**. HaploHiC is available in the GitHub repository (https://github.com/Nobel-Justin/HaploHiC).

Allele-specific Hi-C: A hierarchy of organization

To learn more about the differences between the maternal and paternal copies of the genome, we analyzed the chromatin contact matrices found through Hi-C and separated by HaploHiC. DNA is often separated into its most dominant conformations, euchromatin and heterochromatin (also known as A and B compartments, respectively), which can be identified through Hi-C. Euchromatin, or loosely packed DNA, is associated with accessible DNA and therefore the genes in these regions are often expressed. Heterochromatin, or tightly packed DNA, often suppresses the expression of genes due to dense packing which hinders transcription. A common technique to separate these two groups is to find the first principal component (PC1) of the Hi-C matrix (*31*). It has also been shown that another special vector, called the Fiedler vector, is highly correlated with PC1 and is able to accurately separate these chromatin domains (*32*).

In studies comparing multiple types of cells or cells going through differentiation, areas of

heterochromatin and euchromatin switch corresponding to genes that are activated/deactivated for the specific cell type (*33*). That is, areas that are loosely packed and exhibit gene expression in one setting, become more tightly packed and suppress gene expression in the other setting (and vice versa). We look to explore this phenomenon in the context of the maternal and paternal Hi-C matrices to determine if the two genomes had differing chromatin compartments. By examining the Fiedler vectors for each chromosome through each phase at a 1 Mb scale, we found that there were changes in chromatin compartments for all chromosomes. Interestingly, we noticed that the vast majority of these changes took place on the borders between compartments (instead of an entire region switching compartments). This implies that although the overall architecture of the maternal and paternal genomes are not exactly the same, they generally have similar structures (with the exception of the X chromosome).

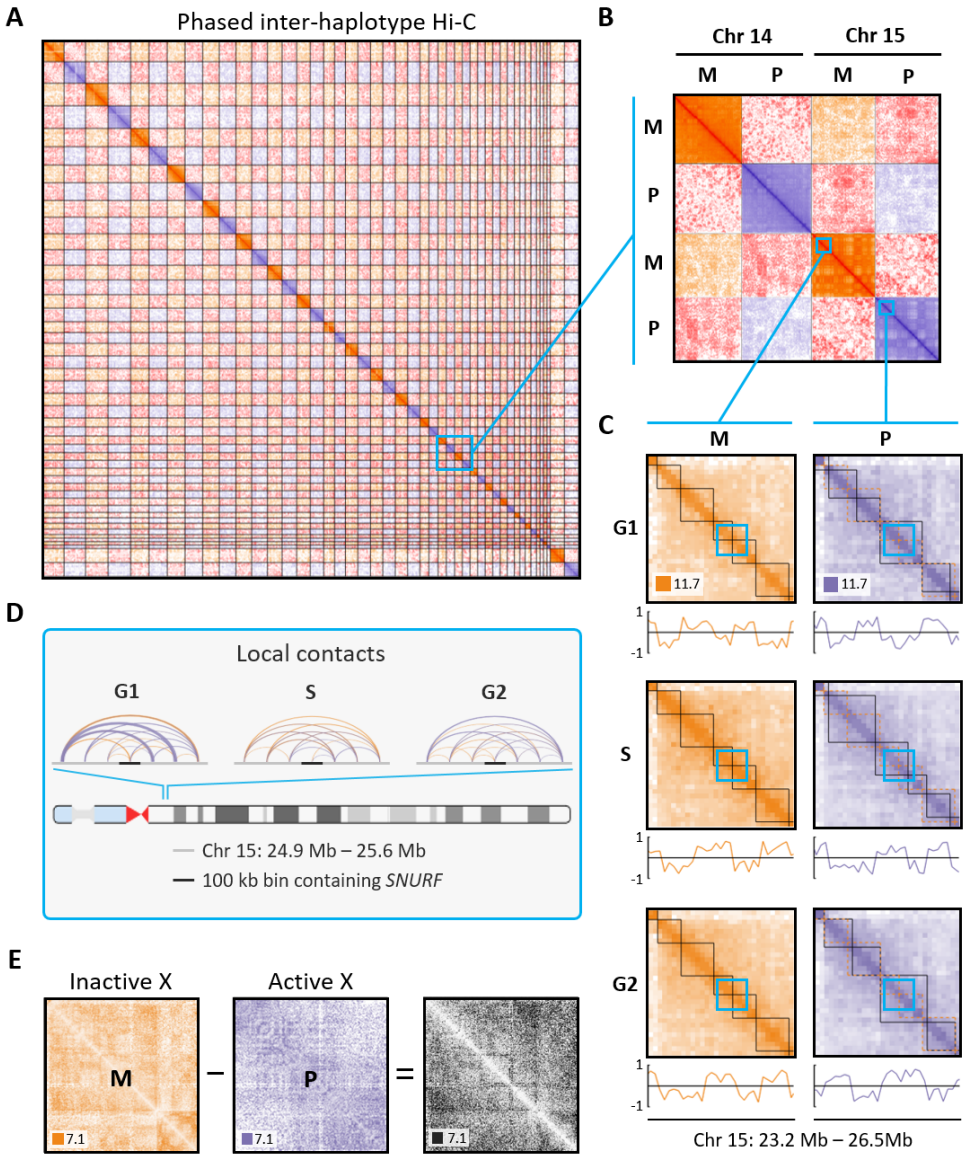

Fig. 6: Allele-specific Hi-C on multiple scales. **(A)** Inter- and intra-haplotype chromatin contacts, shown in log scale 1 Mb resolution (KR normalized) as an average over G1, S, and G2. **(B)** Highlighted region from **A** illustrating inter- and intra-haplotype and chromosomal contacts. **(C)** Subsection of chromosome 15, shown in log scale 100 kb resolution (KR normalized) and separated into parental and cell cycle specific matrices, with parental and cell cycle specific TADs depicted with solid lines. Maternal TADs are preserved as orange dashed lines in paternal matrices, highlighting changes between alleles. **(D)** Representative example of local Hi-C architecture changes, centered around the imprinted gene *SNURF*. **(E)** The inactive maternal X chromosome is partitioned into two superdomains as observed in (*4*), shown in log-scale 100 kb resolution in G1.

Additionally, the Fiedler vector can be calculated recursively to identify Topologically Associated Domains (TADs), as shown by *Chen et al.* (*32*). We applied this method to our Hi-C data for all chromosomes on a 100 kb scale to determine how chromatin architecture changes between the parental genomes and through the cell cycle. In **Figure 6C**, we show an example region within chromosome 15 which highlights the TAD changes between cell cycle phases and parental origin (maternal TADs displayed in paternal matrices as orange dashed lines). TAD boundaries differ when comparing maternal to paternal in each cell cycle phase, as well as between cell cycle phases. These differences appear in all chromosomes and support allelic differences of TADs found in single cells (*34*). Also in **Figure 6C**, we highlight an area in blue signifying the local region surrounding an imprinted gene, *SNURF*, which is expanded in **Figure 6D**. We found that the local genome architecture around *SNURF* had large differences between the maternal and paternal genomes. This was not surprising, as the maternal copy of *SNURF* is completely silenced. In fact, large changes in local chromatin architecture were found for all six imprinted genes.

While the current understanding of genomic architecture dictates that TAD boundaries are invariant (between alleles, cell types, etc), it is also known that intra-TAD structures are highly variable (*33, 34*). Since the spectral method (*32*) has stricter criteria for identifying TADs, and is therefore more sensitive to local structural changes, it often identifies what many call 'sub-TADs' as their own distinct domains. We find that these identified domains are not consistent between alleles or cell cycle phases, and we predict that they are even more highly variable between cell types.

After observing structural differences surrounding genes with large changes in expression like *SNURF*, we look to all genes with differential expression between parents and cell cycle phases. 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 for each differentially expressed gene (700 kb). These matrices represent the local genomic architecture of the differ-

entially expressed genes. To determine whether these local regions have changed in the seven comparisons, we used a method developed by Larntz and Perlman (*35*) which we refer to as the Larntz-Perlman (LP) procedure. This method determines the similarity of two correlation matrices, and outputs whether or not the matrices have statistically significant differences. To satisfy the conditions of the LP procedure, we used the correlation matrix of the log_2-transformed Hi-C data (*35*). We applied the LP procedure to all genes that were differentially expressed in RNA-seq (Bru-seq) and found that 627 (446) genes had at least one transition in which the expression and local architecture had both significantly changed.

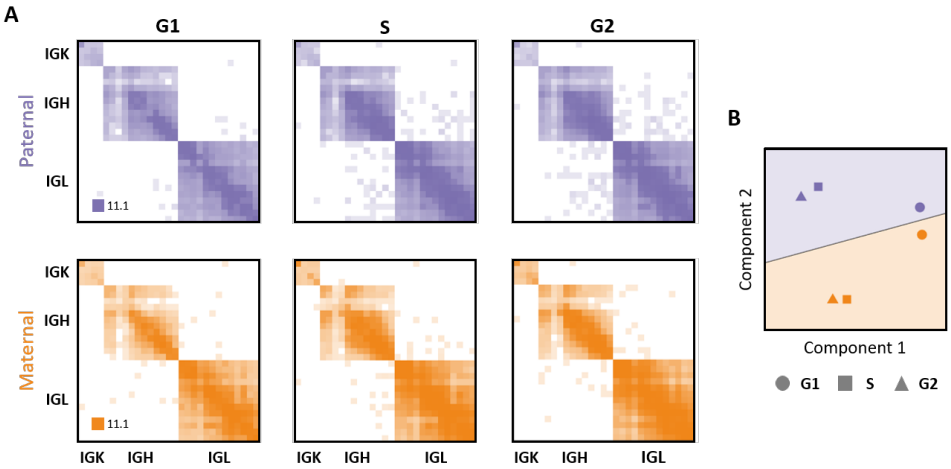

Fig. 7: Allele-specific contacts of immunoglobulin loci. (**A**) Intra- and inter-chromosomal contacts between the *IGK* (chr 2), *IGH* (chr 14), and *IGL* (chr 22) loci in log-scale 1 Mb resolution (raw contacts). Statistically significant differences were found between the maternal and paternal matrices in all three cell phases, and for the paternal matrices between cell cycle phases. (**B**) Maternal and paternal IG loci are easily separated after dimension reduction.

In addition to imprinted and X-inactivated genes, the immunoglobulin (IG) heavy and light chain genes also have known allelic differences in B cells. The B-lymphocyte NA12878 cell line allows us to analyze the IG loci in an allele-specific manner. While maturing, these regions undergo V(D)J recombination which creates a unique pairing IG chain. This chain is then used to assemble B cell receptors (BCR) expressed on the surface of the cell. The generation of functional IG genes during V(D)J recombination is allelicly excluded, meaning that only the

maternal or paternal IG genes will be able to encode the BCR for a particular cell. These genes differ from imprinted and X-inactivated genes in that both parents' alleles can be expressed, but only one parent's IG genes are functional (*36*). To better understand the differences of these genes, we look to chromatin architecture of the IG loci and the contacts between them. Using the LP procedure, we found that the maternal and paternal IG matrices were significantly different in all three cell cycle phases, and that the paternal IG matrices were significantly different between cell cycle phases (**Figure 7**). The maternal and paternal IG matrices are also easily separable after dimension reduction (*37*).

Allele-specific binding

To further investigate the relationship between allele-specific gene expression and architecture, we look to DNA binding proteins such as RNA polymerase II (Pol II), CTCF, and 35 other transcription factors. We used publicly available data from AlleleDB (*8*) in tandem with our allele-specific mature RNA expression data and found that 120 genes have a parental bias in expression and in TF binding. For Pol II, 13 out of 18 cases match (72.2%) in terms of gene expression bias and binding bias. For CTCF, 45 out of 92 cases match (48.9%) and for all other TFs analyzed, 20 out of 30 cases match (66.7%) (**Supplementary Table 6**). The CTCF binding bias of around 50% is expected and has been found in previous studies (*9*).

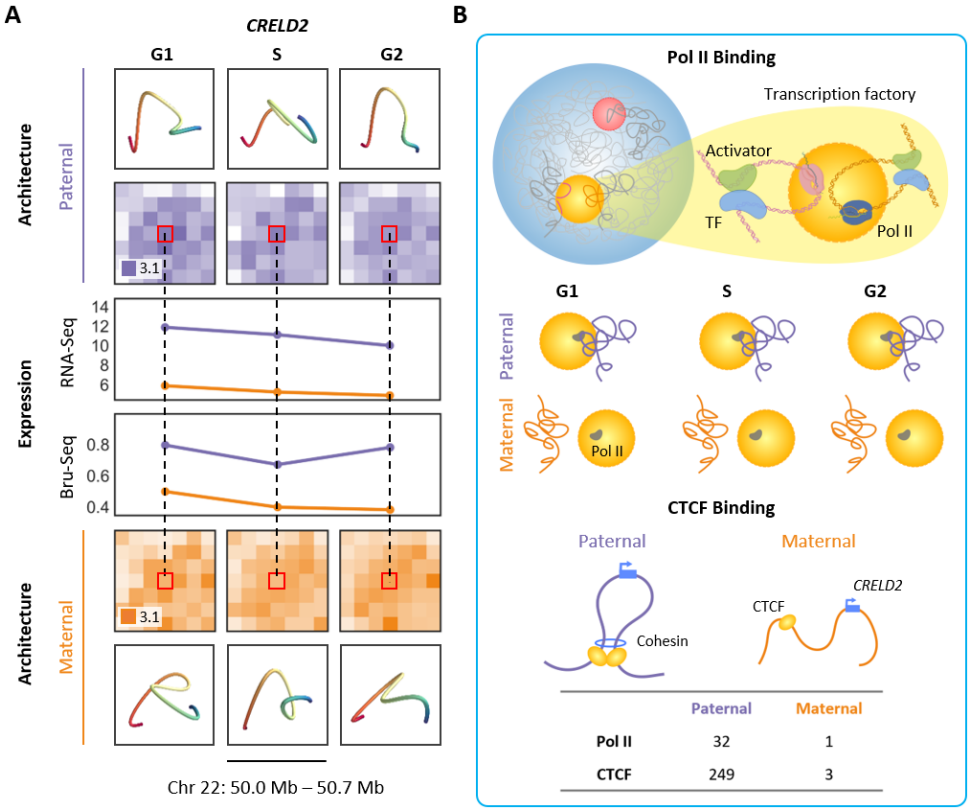

Fig. 8: Allele-specific dynamics of *CRELD2*. **(A)** *CRELD2* has an expression bias towards the paternal allele in all cell cycle phases in mature RNA, and higher expression of paternal nascent RNA. It also has significant changes in local genome architecture, a bias in CTCF binding, and exhibits paternally biased binding of Pol II. Gene expression is quantified by RNA-Seq and Bru-Seq (RPKM) and genome architecture changes are determined using the 700 kb region surrounding the TSS (highlighted by red square) of *CRELD2*. **(B)** Schematic representation of allele-specific Pol II and CTCF binding near the *CRELD2* gene. Bottom table shows counts of Pol II and CTCF binding on SNVs/InDels.

We also analyzed the 18 genes with biases in expression and Pol II binding to see if they had significant changes in local genome architecture. Through the LP procedure, we found that 13 out of the 18 genes had significant changes in structure in at least one cell cycle phase. This lends credence to the idea that changes in structure often coincide with changes in expression, due to the increased or decreased accessibility of a gene for the necessary transcriptional machinery. We highlight the gene *CRELD2*, which exhibits differential expression, differential conformation, and significant biases in binding of both Pol II and CTCF **Figure 8**. The large

paternal bias in Pol II supports the paternal bias of nascent and mature RNA, and the bias in CTCF supports the significant changes in local chromatin architecture around *CRELD2*. The differences in the alleles' local chromatin architecture may induce changes in their ability to access transcription factories (*38*) (**Figure 8B**).

Allele-specific 4D Nucleome analysis

The goal of the allele-specific 4D Nucleome is to integrate genome architecture with gene expression through time, in an allele-specific manner. In order to achieve this goal, one needs to establish a metric to represent structural characteristics which are compatible with functional measures. Hi-C data 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. One way to think about this 3D structure of the genome is through the perspective of a 'genomic network'. In such a network, the genomic regions are considered nodes and the physical contacts between genomic regions are the edges. By thinking of the genome as a network, it opens up new methods of analysis and techniques to learn about the interactions between genomic regions. One concept that is well known in network theory is centrality. Different forms of centrality measurements, called features, encompass a wide range of network properties. The common goal of these features is to assign a quantitative measurement to the structure and organization of the network.

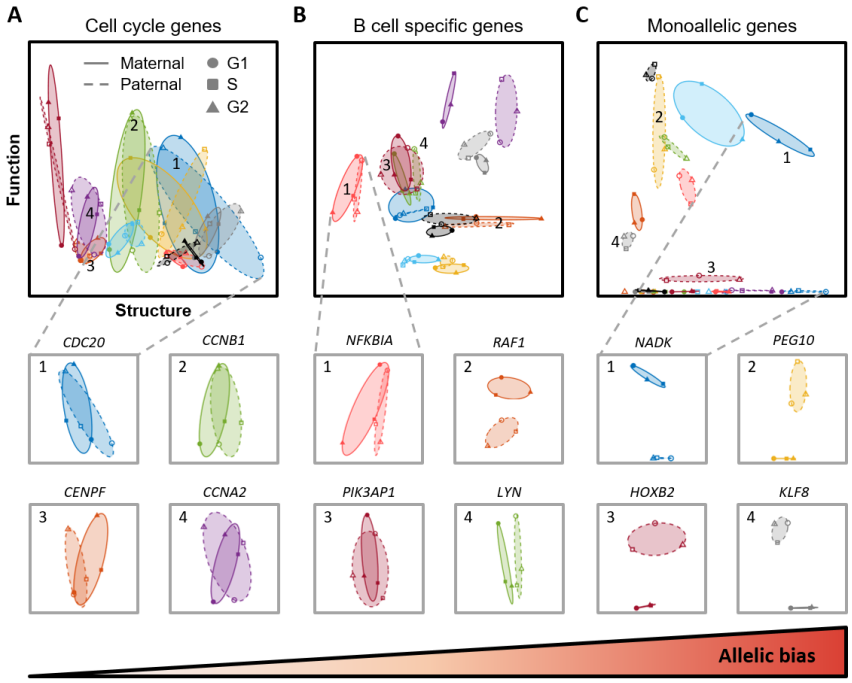

Fig. 9: Allele-specific 4D Nucleome analysis captures variability in allelic bias. Structure-function phase portraits of three sets of genes obtained using an allele-specific 4DN analysis. (**A**) Cell cycle genes have only subtle differences between their maternal and paternal alleles which leads to high overlap in their phase portraits. **(B)** Genes specific to B cell functions have a mix of allelic difference levels. Some B cell genes have a high similarity between alleles, while others have little to no overlap in their phase portraits. **(C)** Monoallelicly expressed genes, including imprinted and X-linked genes, show no allelic overlap in their phase portraits.

Eigenvector centrality has different values assigned to each node in a network, which correspond to that node's influence on the network. Using this feature as a measurement of structure and gene expression for function, we construct a visual representation of the allele-specific 4D Nucleome (**Figure 9**). We call this visualization a phase portrait. The phase portrait allows us to analyze multiple genes, as well as the two alleles of each gene, through the cell cycle. Genes which are known to be crucial for cell cycle regulation have very similar phase portraits between the two alleles, shown by the large overlap between the ellipses that encompass the three cell cycle phases (**Figure 9A**). This shows a robust coordination between alleles to maintain proper cell cycle function. Similar results were found for essential core circadian genes (portraits not shown). Genes specific to B cell identity, have a variety of allelic differences (**Figure 9B**).

The alleles of *PIK3AP1* have very similar phase portraits while *RAF1*'s alleles are completely separated. We hypothesize that the phase portraits of alleles from cell type defining genes will become more similar as the cell specializes. For example, B cell genes may have more similar phase portraits between alleles after the B cell becomes activated and differentiates into a plasma B cell, but further work is required to better understand the relationship between allelic differences of structure and function through cell differentiation. Monoallelicly expressed genes have no overlap in their phase portraits due to their large structural and functional differences (**Figure 9C**). Through the use of phase portraits, we observe significant allelic bias in cell identity genes, and notably absent bias in universally essential cell cycle and circadian rhythm genes. We offer a model in which coordinated biallelic expression reflects prioritized preservation of essential gene sets.

Discussion

In this study, we elucidate the intimate relationship between gene expression, genome architecture, and protein binding in an allele-specific manner throughout the cell cycle. We confirmed known allele-specific properties such as the differential expression of imprinted (*15, 19–21*) and X-linked genes, broad similarities of the A and B compartments between the maternal and paternal genomes, and superdomains of the X chromosome (*4*). Unique to this study, we established a coordination of allele-specific biases between gene expression and protein binding, with corresponding changes in local genome architecture in hundreds of genes not commonly associated with allelic bias. We verified these results and our methods by finding similar patterns in known imprinted and monoallelicly expressed genes (*4*). Through the integration of these structural and functional data through time, we created an allele-specific extension of the 4D Nucleome.

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, many of which are not well appreciated for their allele-specific differ-

ences. Approximately half of the genes with ABE are only differentially expressed in certain cell cycle phases, and many genes with CBE only show differential expression in one allele or the other. Further research is needed to explore why certain genes have coordinated cell cycle dynamics in their alleles, while other genes have disparate expression in some cell cycle phases.

Since allele-specific Hi-C contact matrices are innately more sparse than unphased Hi-C contact matrices, we sought a way to phase additional loci. We created a novel phasing algorithm, HaploHiC, which uses reads mapped to SNVs/InDels, to predict nearby reads of unknown origin. This allowed to us to decrease the sparsity of our allele-specific contact matrices, and therefore have more confidence in our analysis of the parental differences in genome architecture. While we found that the overall organization (A/B compartments) of the two genomes are broadly similar, there are many differences in TADs and local genome architecture. These significant differences were observed both between alleles and in an allele-specific manner between cell cycle phases.

We focused our search of allele-specific differences in genome architecture by calculating the similarity of local contacts surrounding the TSS of differentially expressed genes through the Larntz-Perlman procedure. We found that the majority of genes with differential expression, either between alleles or between cell cycle phases, had corresponding changes in local architecture in at least one comparison. This finding was supported by narrowing our analysis on genes that were monoallelically expressed through X-inactivation, imprinting, or other means. These monoallelically expressed genes were found to have 80.3% of their differentially expressed comparisons (maternal vs paternal in G1, S, and G2) match with corresponding changes in local genome architecture. This further supports the intimate relationship between allele-specific genome structure and gene expression. Through the use of a fully haplotyped B-lymphocyte cell line, we were also able to observe the allelic exclusion of IG loci. We analyzed the IG loci for allelic differences in genome architecture and found that the submatrix containing the IGK, IGH, and IGL loci was significantly different between the maternal and paternal copies.

In order to explore the relationship between allele-specific gene expression and genome architecture, we incorporated publicly available allele-specific binding data (*9*) for a number of transcription factors, including Pol II and CTCF. In genes that had both expression and Pol II binding biases, we found that 72% of these genes had matching allele-specific bias. Additionally, we found that 11 of the 13 matching genes had significant changes in local genome architecture as determined through the LP procedure. These data support an intimate relationship between genome structure and function, through the binding of transcription factors. 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 affected 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 creation of an allele-specific 4DN, which will be imperative to discern allelic differences underlying many immune-related diseases and cancers in future studies.

1 Materials and Methods

We grew the NA12878 cells in RPMI1640 medium supplemented with 10% fetal bovine serum (FBS). Live cells were stained with 10um of Hoechst 33342 (Cat #B2261, Sigma-Aldrich), and then subjected to Fluorescence-Activated Cell Sorting (FACS) to obtain cell fractions at the corresponding cell cycle phases G1, S, and G2/M. Total RNA were extracted from sorted live cells for RNA-seq and Bru-seq. For cells used in Hi-C libraries construction, we cross-linked the cells with 1% formaldehyde, neutralized the cross-linking reaction with 0.125M of glycine. The cells were stained with Hoechst and subjected to FACS to isolate cells at cell cycle stages G1, S, and G2/M.

1.1 Hi-C library construction

Hi-C library construction and Illumina sequencing was done using established methods (*4*). 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 University of Michigan DNA Sequencing Core facilities (Seq-Core).

1.2 Alignment and filtration of Hi-C PE-reads

Illumina adapter sequences and low quality ends were trimmed from raw Hi-C reads by Trimmomatic (*39*). Paired-ends (R1 and R2) were separately aligned to the human reference genome (hg19) using BWA (0.7.15) (*40*), and realigned by GATK (3.5) (*41*). The VCF file recording phased germline mutations of NA12878 was downloaded from AlleleSeq database (version: Jan-7-2017) (*9*), recording a total of 2.49 million heterozygous genomic loci that have phased

germline SNVs/InDels (**Supplementary Table 7**). PE-reads were filtered out if either or both ends are un-mapped, mapped with mapping quality less than 20, or mapped to multiple genomic positions. Note that supplementary alignments were reserved. After the exclusion of invalid pairs, remaining Hi-C PE-reads were eligible to provide chromatin contacts (**Supplementary Table 9, Supplementary Notes 2.1.2 and 2.1.3**).

1.3 Category of Hi-C PE-reads based on parental origin

Hi-C reads are marked as haplotype-known or -unknown depending on coverage of heterozygous SNVs and InDels (**Supplementary Notes 2.1.1**). If sequencing reads cover a phased heterozygous genomic position and match either of parental alleles, it is considered as haplotype-known and marked with corresponding haplotype (i.e., 'P' for paternal or 'M' for maternal), otherwise it will be considered as haplotype-unknown (shortened as 'U'). Note that one read might support both haplotypes simultaneously (marked with 'I' as inter-haplotype), as it covers the junction site of the ligated paternal and maternal fragments. Parental alleles from the haplotype-known reads are required to have a base quality of at least 20 and be no less than 5 bp away from the bilateral read edges.

According to haplotype status of paired-ends, Hi-C PE-reads are assigned to seven categories: dEnd-P/M/I, sEnd-P/M/I, and dEnd-U (**Figure 5A, Supplementary Table 8, Supplementary Notes 2.1.4**). Here, 'dEnd' and 'sEnd' represent dual-ends and single-end, respectively. Four categories (dEnd-P/M/I and sEnd-I) have confirmed allele specific contacts, and are defined as phased. Conversely, the other three categories are unphased; their PE-reads have one (sEnd-P/M) or two (dEnd-U) haplotype-unknown reads.

1.4 Assign haplotype to unphased Hi-C PE-reads

We propose a local contacts based algorithm to impute the haplotype for haplotype-unknown reads (**Figure 5B, Supplementary Figure 1, Supplementary Notes 2.1.5-2.1.7**). For each read of unphased PE-reads, the local region is obtained by bilateral extension (50 kb unilater-

ally) from the mapped position. Each unilateral region might be trimmed to include at most 30 phased heterozygous loci. Based on Hi-C pairs from phased categories, four kinds of haplotype combination contacts (P-P, M-M, P-M, and M-P) are calculated for paired local regions of unphased PE-reads. One haplotype combination is assigned to unphased PE-reads randomly with a probability relative to the known haplotype contacts count. If unphased PE-reads belong to sEnd-P/M categories, haplotype combinations lacking the existing haplotype will be excluded. If all haplotype combination contacts are zero, paired local regions will be further extended, step-wise, until this region has non-zero contacts or the regions reach the maximum length (1 Mb).

1.5 Contacts matrix

After haplotype assignment, Hi-C PE-reads are distributed to intra-haploptye (P-P and M-M) and inter-haplotype (P-M and M-P). Juicer (*42*) is applied on intra-haploptye PE-reads, and outputs paternal and maternal contacts matrices, respectively. Inter-haplotype contact matrices are generated by HaploHiC (**Supplementary Notes 2.1.8**). 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. Matrices are normalized through the Knight-Ruiz method of matrix balancing (*43*) and through Toeplitz normalization (*32*) unless otherwise specified.

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 Mats Ljungman for generating the Bru-seq sequencing libraries and for helpful discussion. We also thank Dr. Jacob Kitzman for providing the NA12878 cell line and assisting in resolving genotype-phasing issues of the cell line. We are grateful to Gabrielle Dotson and 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 request should be addressed to I.R.: indikar@umich.edu.

References

1. David U Gorkin, Yunjiang Qiu, Ming Hu, Kipper Fletez-Brant, Tristin Liu, Anthony D Schmitt, Amina Noor, Joshua Chiou, Kyle J Gaulton, Jonathan Sebat, et al. Common dna sequence variation influences 3-dimensional conformation of the human genome. *Genome biology*, 20(1):1–25, 2019.

2. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

3. Alexander Gimelbrant, John N Hutchinson, Benjamin R Thompson, and Andrew Chess. Widespread monoallelic expression on human autosomes. *Science*, 318(5853):1136–1140, 2007.

4. 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.

5. Chloe M Rivera and Bing Ren. Mapping human epigenomes. *Cell*, 155(1):39–55, 2013.

6. Peter H Sudmant, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75, 2015.

7. Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

8. 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.

9. 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.

10. William W Greenwald, He Li, Paola Benaglio, David Jakubosky, Hiroko Matsui, Anthony Schmitt, Siddarth Selvaraj, Matteo DAntonio, Agnieszka DAntonio-Chronowska, Erin N Smith, et al. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nature communications*, 10(1):1–17, 2019.

11. Virginia Savova, Sung Chun, Mashaal Sohail, Ruth B McCole, Robert Witwicki, Lisa Gai, Tobias L Lenz, C-ting Wu, Shamil R Sunyaev, and Alexander A Gimelbrant. Genes with monoallelic expression contribute disproportionately to genetic diversity in humans. *Nature genetics*, 48(3):231, 2016.

12. Anwesha Nag, Virginia Savova, Ho-Lim Fung, Alexander Miron, Guo-Cheng Yuan, Kun Zhang, and Alexander A Gimelbrant. Chromatin signature of widespread monoallelic expression. *Elife*, 2:e01256, 2013.

13. 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.

14. Huntington F Willard. X chromosome inactivation, xist, and pursuit of the x-inactivation center. *Cell*, 86(1):5–7, 1996.

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

16. Christelle Borel, Pedro G Ferreira, Federico Santoni, Olivier Delaneau, Alexandre Fort, Konstantin Y Popadin, Marco Garieri, Emilie Falconnet, Pascale Ribaux, Michel Guipponi, et al. Biased allelic expression in human primary fibroblast single cells. *The American Journal of Human Genetics*, 96(1):70–80, 2015.

17. CL Clarke, J Sandle, AA Jones, A Sofronis, NR Patani, and SR Lakhani. Mapping loss of heterozygosity in normal human breast cells from brca1/2 carriers. *British journal of cancer*, 95(4):515, 2006.

18. Bernard Friedenson. The brca1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers. *BMC cancer*, 7(1):152, 2007.

19. 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.

20. 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*, pages gr–192278, 2015.

21. 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.

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

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

24. 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.

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

26. 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.

27. Juan Carlos Rivera-Mulia, Andrew Dimond, Daniel Vera, Claudia Trevilla-Garcia, Takayo Sasaki, Jared Zimmerman, Catherine Dupont, Joost Gribnau, Peter Fraser, and David M Gilbert. Allele-specific control of replication timing and genome organization during development. *Genome research*, 28(6):800–811, 2018.

28. Peiyao A Zhao, Takayo Sasaki, and David M Gilbert. High-resolution repli-seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome biology*, 21(1):1–20, 2020.

29. 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.

30. 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.

31. 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.

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

33. 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.

34. 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.

35. Scott Ronquist, Sijia Liu, Michael Perlman, and Indika Rajapakse. 4dnvestigator: a toolbox for the analysis of time series hi-c and rna-seq data. *Nucleic Acids Research*, 2019 Submitted.

36. Christian Vettermann and Mark S Schlissel. Allelic exclusion of immunoglobulin genes: models and mechanisms. *Immunological reviews*, 237(1):22–42, 2010.

37. Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

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

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

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

41. 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*, 2010.

42. 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.

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

44. 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.

45. 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.

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

47. 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.

48. 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.

49. 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.

50. 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.

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

52. 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.

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

2 Supplementary Notes

2.1 Haplotype assignment of Hi-C PE-reads

2.1.1 Phased germline mutations of NA12878

From AlleleSeq database (version: Jan-7-2017) (*9*), we downloaded the VCF file of sample NA12878. This VCF file contains phased heterozygous germline mutations (SNV and InDel) with a total of 2.49 million genomic loci (**Supplementary Table 7**). 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 example, 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 7**). HaploHiC outputs a list of these filtered genotypes. The minimum distance is also applied in InDel allele judgment on sequencing reads (see 2.1.4).

2.1.2 Alignment of Hi-C PE-reads

After trimming adapter sequences and low quality ends by Trimmomatic (*39*), paired-ends (R1 and R2) were separately aligned to human reference genome (hg19) using the function 'mem' from BWA (0.7.15) (*40*) with default parameters. Then, each alignment BAM file was coordinated, sorted, and realigned by GATK (3.5) (*41*) with the VCF file mentioned above (section 2.1.1). All realigned BAM files were sorted by read names before processing in HaploHiC.

2.1.3 Unqualified and invalid aligned Hi-C PE-reads

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 8**).

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 (*44*), 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).

2.1.4 Parental origin categories of Hi-C PE-reads

HaploHiC assigns haplotype situation to sequencing reads by checking whether reads covers and supports 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 5**): 1) 'P' (only paternal allele), 2) 'M' (only maternal allele), 3) 'I' (both paternal and maternal 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 paternal and maternal 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 2.1.1) 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 NA12878 (**Supplementary Table 7**).

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 5**. Additionally, we also set requirements on distance of alignments (**Supplementary Table 8**).

- 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 paternal and maternal 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 paternal and maternal 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.

2.1.5 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 5B**). 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.

2.1.6 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.

2.1.7 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 5B**), 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.

2.1.8 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 9**). 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 9**), which supports the accuracy of HaploHiC phasing.

2.1.9 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 (*45*) to construct pairwise allele specific contacts of paternal and maternal genomes respectively (**Supplementary Table 10**). The paternal and maternal genomes are downloaded from AlleleSeq database (version: Jan-7-2017) (*9*). The simulations on the LOOP level are CN-based (copy number), and the others are SV-based (structure variation, **Supplementary Figure 2**). 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 2A**). 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 2B**). 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 (*46*) 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 10**). 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' contact, 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).

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

The live cells sorted were used for total RNA extraction. Subsequently RNA-seq library construction was carried out in the seq-core facility, and sequence reads of 50-base in length were generated on an Illumina HiSeq 2500 station.

2.3 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 subjected to FACS 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 (*31*). 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.

2.4 Allele specific RNA-seq and Bru-seq analyses

The normal RNA-seq and Bru-seq analysis were performed as previously described (*47, 48*). 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 (*48*) for full details). Similarly, in RNA-seq data process, 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/Cuffdiff (version 2.2.1) was used for expression quantification and differential expression analysis, 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 (*49*). 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 (*50*). 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. The statistical significance of allele specific expression for each SNV was then quantified by a binomial test with a null probability of 0.5.

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. Paternal and maternal abundance of each gene were calculated by multiplying the overall abundance estimate by the fraction of the SNV-covering reads that were paternal and maternal, respectively.

Gene level significance of allele specific expression for each gene in each cell cycle stage was evaluated using a negative binomial model variance estimation and is improved through a local regression relating variance to the mean (https://www.mathworks.com/help/bioinfo/ref/nbintest.html). The same method was used to determine differential gene expression between the overall abundance in different cell cycle stages.

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 5,003 genes with allele read counts $\geq$5, which is the minimum number of counts that we use to reliably estimate ABE. We identified differentially expressed genes between alleles and between cell cycle phases using a MATLAB implementation of DESeq, which uses a negative binomial model to determine if two groups of samples are statistically significantly different in gene expression (https://www.mathworks.com/help/bioinfo/ug/identifying-differentially-expressed-genes-from-rna-seq-data.html) (*25*). We imposed a minimum RPKM level of 0.01, false discovery rate adjusted p-value threshold of 0.05, and a fold change cutoff of FC $>$ 2 (*26*). We identified 778 genes (FDR $<$ 0.05) which were differentially expressed, 605 genes with allele biased expression and 284 genes with cell cycle biased expression (111 genes in common).

Bru-seq detects nascent transcripts containing both exons and introns. Sequence reads from both exons and introns containing informative SNVs are used to evaluate ABE. 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 show low read coverage depth to be able to be statistically evaluated compared to that in the RNA-seq data. We require allelic count $\geq$ 5 to estimate ABE 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, and we found that 556 genes were differentially expressed, 438 genes with allele biased expression and 171 genes with cell cycle biased expression (53 genes in common).

2.5 The Laplacian: Fiedler number and Fiedler vector

The Laplacian, Fiedler number, 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 number, and the corresponding eigenvector is called the Fiedler vector (*51*). In the context of the 4DN, the magnitude of the Fiedler number is a measure of the underlying stability of the topology of the genomic region at any given scale. For Hi-C contacts, a high Fiedler number in a genomic region would suggest high conformational stability that may be important for regulation of gene expression. 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 compo-

nent 1) of the correlation matrix of the normalized Hi-C matrix (*31*), but it is directly related to properties of the associated graph (*51*). The Fiedler vector can be calculated recursively to identify smaller partitions in Hi-C data. These smaller partitions correspond to topologically associated domains (TADs) (*32*). Description of the Laplacian, Fiedler number, and Fiedler vector adapted from (*24*).

2.6 Larntz-Perlman procedure

Algorithm 1: LP method

Input: Hi-C matrices $\mathbf{A}^{(m)} \in \mathbb{R}^{n\times n}$, $m = 1, \ldots, k$

Output: *P*-value p, and test statistic $\mathbf{S}$

1 Compute the correlation matrix $\mathbf{C}^{(m)} = \text{corr}(\log_2(\mathbf{A}^{(m)}))$. Define the corresponding population correlation matrices as $\mathbf{P}^{(m)}$

2 Define the null hypothesis
$$H_0 : \mathbf{P}^{(1)} = \cdots = \mathbf{P}^{(k)}$$

3 Compute the Fisher z-transformation $\mathbf{Z}^{(m)}$. Elements in $\mathbf{Z}^{(m)}$ and $\mathbf{C}^{(m)}$ are denoted as $z_{ij}^{(m)}$ and $c_{ij}^{(m)}$, respectively, and
$$z_{ij}^{(m)} = \frac{1}{2}\ln\left[\frac{1+c_{ij}^{(m)}}{1-c_{ij}^{(m)}}\right]$$

4 Form $\mathbf{S}$, where elements in $\mathbf{S}$ are denoted as s_{ij} and
$$s_{ij} = (n-3)\sum_{m=1}^{k}(z_{ij}^{(m)} - \bar{z}_{ij})^2,\ \bar{z}_{ij} = k^{-1}\sum_{m=1}^{k} z_{ij}^{(m)}$$

5 Calculate the test statistic
$$T = \max_{1\le i<j\le n} s_{ij}$$

6 Reject H_0 at level α if $T > \chi^2_{k-1,\epsilon(\alpha)}$, where $\chi^2_{k-1,\epsilon(\alpha)}$ is the chi-squared distribution with $k-1$ degrees of freedom and $\epsilon(\alpha) = (1-\alpha)^{2/n(n-1)}$ is the Šidák correction

7 Calculate p, the level α at which $T > \chi^2_{k-1,\epsilon(\alpha)}$

Return: p and $\mathbf{S}$

Algorithm 1 is adapted from (*35*).

2.7 Structure alignment and visualization

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 (*52*) which applies the optimal transform to the second matrix (including scaling/dilation, rotations, and reflections) to minimize the sum of square errors of the point-wise 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 (*53*).

3 Supplementary Figures

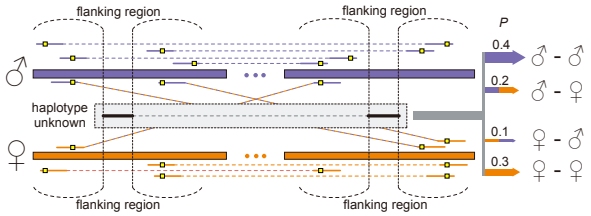

Fig. 1: Expanded diagram of HaploHiC phasing in Figure 5. Haplotype-unknown Hi-C pair is distributed to certain haplotype combination with probabilities depending on local phased contacts in flanking regions. Homologous genomic regions are linked by phased Hi-C pairs (purple for paternal, orange for maternal). Phased heterozygous loci are denoted by yellow squares.

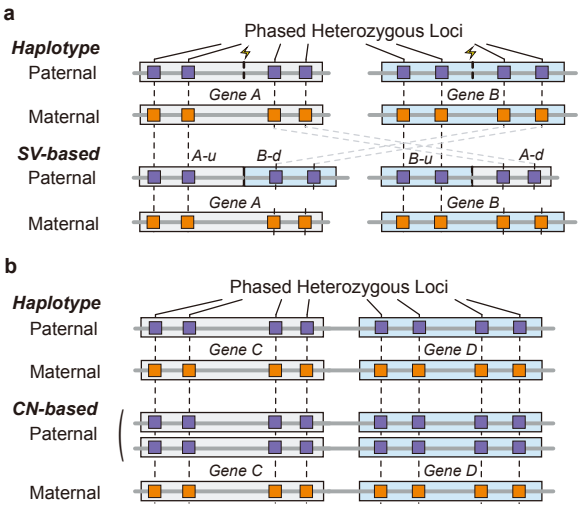

Fig. 2: Simulation of SV-based and CN-based contacts. **(A)** Gene pairs are reciprocally partially concatenated on the paternal haplotype to form a chimeric sequence. The maternal haplotype carries the wide-type, and 'u' and 'd' denote upstream and downstream genomic regions, respectively. **(B)** The copy number of gene pairs' genomic region on the paternal haplotype is increased, and the maternal haplotype is kept at one copy.

4 Supplementary Tables

Supplementary Table 1: Differential expression per chromosome. Number of allele specific genes (AS), differentially expressed genes (DE), and percentage of AS 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 with matching maternal (orange) and paternal (purple) gene expression and TF binding biases. Genes with conflicting parental biases of expression and TF binding are left unshaded.

Supplementary Table 7: Details of phased germline mutations of NA12878. Mutation counts of diverse allele scenarios on paternal and maternal genomes are listed.

Supplementary Table 8: 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 9: Details of Hi-C PE-reads in sample NA12878 processed by HaploHiC. Hi-C pair counts of all categories in three steps are listed.

Supplementary Table 10: Details of ten categories of gene pairs in Hi-C simulation and evaluation of HaploHiC. In total, 67 gene pairs details are listed with parental and maternal origin contacts division results from HaploHiC.