Haplotype-resolved karyotype construction from Hi-C data using refLinker

Gregory J. Brunette[a,b,c,d,2], **Richard W. Tourdot**[b,e], **Darawalee Zong**[f], **David Pellman**[a,c,d,g], **and Cheng-Zhong Zhang**[b,d,e,h 2]

This manuscript was compiled on March 2, 2024

Chromosomal aberrations are prevalent in cancer genomes, yet it remains challenging to resolve the long-range structure of rearranged chromosomes. A key problem is to determine the chromosomal origin of rearranged genomic segments, which requires chromosome-length haplotype information. Here we describe `refLinker`, a new computational method for whole-chromosome haplotype inference using external reference panels and Hi-C. We show that `refLinker` ensures consistent long-range phasing accuracy in both diploid human genomes and aneuploid cancers, including regions with loss-of-heterozygosity and high-level focal amplification. We further demonstrate the feasibility of complex genome reconstruction using haplotype-specific Hi-C contacts, revealing new karyotype features in two widely studied cancer cell lines. Together, these findings provide a new framework for the complete resolution of long-range chromosome structure in complex genomes and highlight the unique advantages of Hi-C data for reconstructing cancer genomes with chromosome-scale continuity.

haplotype phasing | chromosomal rearrangements | genome stability

Introduction

Cytogenetic analysis and DNA sequencing have revealed widespread alterations to chromosome number and structure in cancer genomes, advancing our understanding of how specific chromosomal rearrangements contribute to disease (1–3). Despite these advances, there remains a major gap between nucleotide-level information and the long-range structure of rearranged chromosomes. Shotgun sequencing can resolve DNA breakpoints at base-pair resolution but cannot place these breakpoints in the larger context of rearranged chromosomes. On the other hand, cytogenetic karyotyping enables the direct visualization of large-scale chromosomal aberrations but is unable to precisely map genomic breakpoints. Moreover, neither shotgun sequencing nor cytogenetics can accurately determine the parental chromosome origin of rearranged DNA segments, limiting our capacity to investigate both how these alterations arise and how they mediate allele-specific effects.

Long read sequencing has enabled *de novo* assembly of complete chromosomes in haploid (4) and diploid human genomes (5–7). However, the algorithms for long read assembly are designed for euploid genomes and assume uniform read coverage across the genome. In cancer genomes, the presence of large regions with loss-of-heterozygosity or allelic imbalance can confound both the assembly and the scaffolding of assembled contigs of each haplotype. Moreover, cancer genomes harbor large duplicated segments ($1-10$ Mb) with little genetic variation (<1 variant per 100 kb), making it difficult or impossible to determine the breakpoints joining such segments either directly from long reads ($10-100$ kb) or from assembled contigs.

The difficulties in resolving cancer genomic rearrangements highlight a fundamental problem in resolving the structure of non-haploid genomes: determining the genomic locations of homologous DNA sequences contained in either rearranged chromosomes or different parental homologs. For diploid genomes, this problem can be overcome with the use of Hi-C sequencing, which provides sparse haplotype linkage at the chromosome scale (8). Additionally, the feasibility to achieve similar chromosome-level resolution of aneuploid cancer genomes has been demonstrated in recent studies (9–13). However, because Hi-C only offers sparse linkage between distal variants, all these studies require long read sequencing as the starting point to determine short-range haplotype blocks with dense variant coverage.

Here, we demonstrate a strategy to achieve haplotype-resolved construction of cancer karyotypes using Hi-C data alone. We present `refLinker`, a new computational method that combines statistical phasing and Hi-C linkage to derive whole-chromosome haplotypes with $>99\%$ accuracy and encompassing $>99.9\%$ of common variants. We show that `refLinker` ensures consistent long-range phasing accuracy in aneuploid genomes, even in the presence of allelic imbalance and structural complexity. We then demonstrate the feasibility of karyotype reconstruction of cancer genomes using haplotype-specific Hi-C contacts with examples of large segmental duplications, inversion, and high-level amplification. This work provides a solution for the complete resolution of long-range chromosome structure in complex genomes.

Results

A. Limitations of molecular phasing and statistical phasing. Prior studies have demonstrated both the utility and limitations of Hi-C data for haplotype inference (11, 14–16). Briefly, intrachromosomal Hi-C contacts can provide chromosome-scale haplotype linkage that ensures long-range phasing accuracy, but the sparsity of long-range Hi-C contacts limits the completeness of haplotype inference (14). To overcome this limitation, long reads or linked-reads are used to produce short-range haplotype blocks with dense coverage of heterozygous genotypes, which are then scaffolded by long-range Hi-C contacts (11, 15).

In addition to molecular phasing from long reads, local haplotype linkage between common variants can be inferred from reference haplotypes in a genotyped population by statistical phasing (17, 18). However, haplotype inference from statistical phasing is prone to long-range switching errors due to recombination. To evaluate the long-range accuracy of statistical phasing, we

Author affiliations: [a]Department of Cell Biology, Blavatnik Institute, Harvard Medical School, Boston, MA, USA; [b]Department of Biomedical Informatics, Blavatnik Institute, Harvard Medical School, Boston, MA, USA; [c]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA; [d]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [e]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA; [f]National Cancer Institute, National Institutes of Health, Bethesda, MD, USA; [g]Howard Hughes Medical Institute, Chevy Chase, MD, USA; [h]Department of Pathology, Dana-Farber Cancer Institute, Boston, MA, USA

[2]To whom correspondence should be addressed. E-mail: gbrunette@g.harvard.edu, cheng-zhong_zhang@dfci.harvard.edu

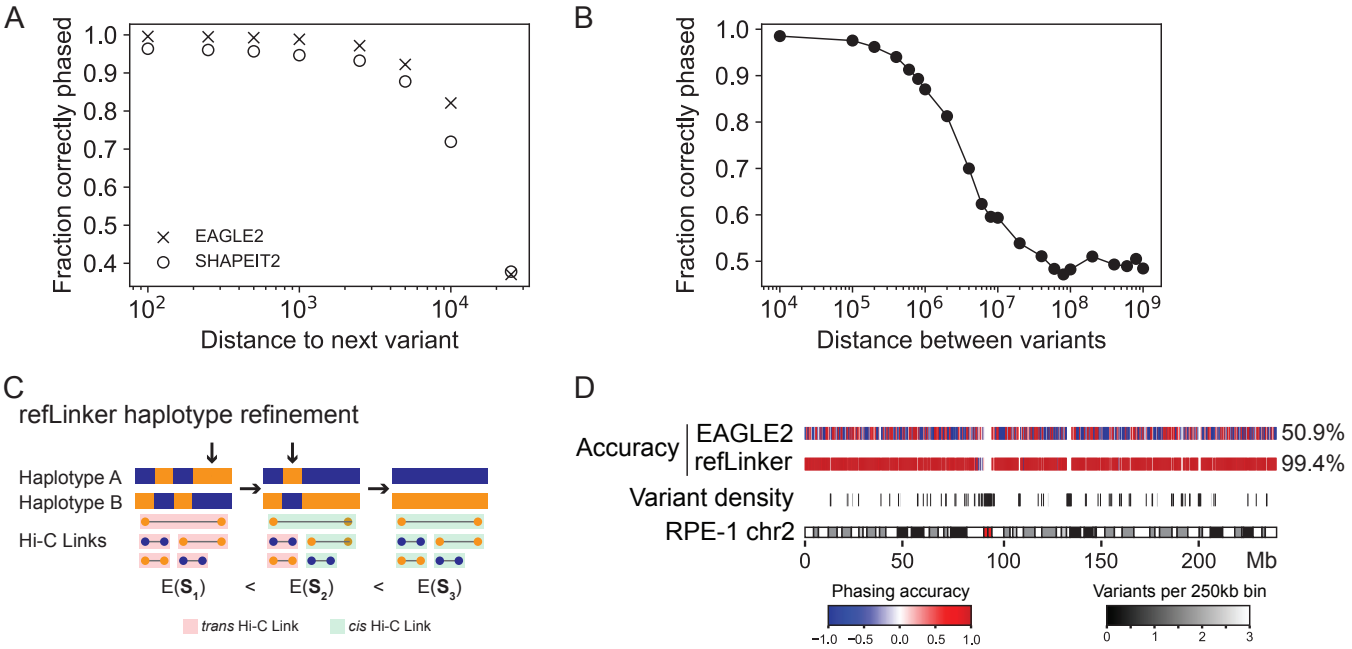

Fig. 1. Determination of whole-chromosome haplotypes from statistical phasing and Hi-C. (A) Fraction of correct haplotype linkage is plotted for adjacent variants on RPE-1 chr2 at different genomic distances for the statistical phasing algorithms EAGLE2 ($\times$) and SHAPEIT2 ($\circ$). (B) Fraction of correct haplotype linkage derived by EAGLE2 for all variants on RPE-1 chr2. For (A) and (B), phasing accuracy was assessed using truth data from experimental phasing and only for single-nucleotide polymorphisms. (C) Schematic of the strategy to correct long-range switching errors by minimizing apparent trans Hi-C contacts. The starting point is the haplotype from statistical phasing (left), with blue and orange representing the true parental haplotype (right). Switching errors are indicated by transitions between blue and orange. Hi-C links are shown below as pairs of blue or orange circles. Hi-C links indicating cis contacts (A-A or B-B) are colored green and trans contacts (A-B) are colored red. refLinker iteratively searches for switching blocks (arrows) to minimize the number of apparent trans Hi-C links. (D) Switching errors in statistical phasing are eliminated by refLinker to generate accurate whole-chromosome haplotypes. Top row, switching errors in the EAGLE2 haplotype are indicated by red-to-blue phasing switches (500 kb bins) on chr2 of the RPE-1 genome; the average phasing accuracy of 50.9% reflects roughly equal contributions of both parental haplotypes in the EAGLE2 solution. Second row, consistent phasing accuracy after corrections using Hi-C links by refLinker. Third row, variant density (250 kb bins) showing gaps in the haplotype solution corresponding to variant-poor regions, including the centromere.

compared results from two well-established methods, EAGLE2 (18) and SHAPEIT2 (17) to truth haplotype data in the RPE-1 cell line that we determined previously from linked-reads and Hi-C (11). The EAGLE2 and SHAPEIT2 statistical haplotypes were generated using reference haplotype data from the 1000 Genomes Project Phase 3 (19) on heterozygous variants in RPE-1. Results generated by both tools show highly accurate linkage between variants within 1 kb ($> 95\%$), with EAGLE2 showing better overall accuracy than SHAPEIT2 (Figure 1A). However, the accuracy drops significantly for variants that are further apart ($70 - 80\%$ at 10 kb). Moreover, the accumulation of switching errors over many variants prevents long-range haplotype inference over intervals > 1 Mb (Figure 1B).

We reasoned that long-range switching errors in statistical phasing haplotypes may be corrected with Hi-C linkage to achieve chromosome-scale accuracy. The feasibility of this strategy is demonstrated by a recent method, IntegratedPhasing, which samples statistical phasing linkage from the SHAPEIT2 haplotype graph for integration with Hi-C reads, achieving accurate long-range haplotype inference in the diploid genome, NA12878 (16). However, because this approach relies on the assumption of uniform sequencing coverage between homologs and across the genome, it remained unclear whether the same strategy could be applied to aneuploid cancers. To overcome this limitation, we developed `refLinker` (https://github.com/gbrunette/refLinker), a new approach for complete haplotype inference from statistical phasing and Hi-C that is directly applicable to aneuploid genomes. We further perform a systematic evaluation of the long-range phasing accuracy of IntegratedPhasing and `refLinker` using Hi-C data of both diploid genomes and aneuploid cancer genomes with deletions, duplications, and higher-level copy-number alterations.

B. A general formalism of integrative haplotype inference. As described in our previous study (11), we use a binary representation for heterozygous genotypes and their corresponding haplotype phase. Briefly, reference and alternate genotypes at a heterozygous position i are represented as $\sigma_i = +1$ or -1, respectively and the chromosomal haplotype is represented as the vector

$$\mathbf{S} = (s_1, s_2, \ldots s_N) \quad s_i = \pm 1. \qquad [1]$$

Molecular linkage from Hi-C reads spanning two or more heterozygous sites is similarly represented as

$$\mathbf{\Sigma} = \{\sigma_i | 1 \leq i \leq N\} \quad \sigma_i = \pm 1. \qquad [2]$$

This formalism reduces haplotype inference to finding $\mathbf{S}$ that maximizes

$$\sum_{k=1}^{M} \sum_{1 \leq i,j \leq N} \sigma_{i,k} \sigma_{j,k} s_i s_j, \qquad [3]$$

where the first sum is carried over all molecular links and the second over all variant sites. The maximization is equivalent to finding the haplotype phase with the lowest number of clashes, i.e.

$$s_i s_j \sigma_{i,k} \sigma_{j,k} = -1.$$

We further introduce the aggregate molecular linkage between two sites i and j as

$$M_{ij} = \chi_{ij} \sum_{k=1}^{m} \sigma_{i,k}\sigma_{j,k}, \quad [4]$$

with a weighting factor

$$\chi_{ij} = \frac{1}{m} \left| \sum_{k=1}^{m} \sigma_{i,k}\sigma_{j,k} \right| \in [0,1] \quad [5]$$

that measures the concordance of links between two variant sites; χ_{ij} serves to penalize linkage evidence from false variant sites with links to both parental haplotypes.

The optimal haplotype solution based on Hi-C links should maximize:

$$E(\mathbf{S}) = \frac{1}{2} \sum_{ij} M_{ij} s_i s_j. \quad [6]$$

Where, for $\chi_{ij} \approx 1$, we have

$$E(\mathbf{S}) \approx (\#\text{ links consistent with } \mathbf{S}) - (\#\text{ links inconsistent with } \mathbf{S}). \quad [7]$$

Let $\mathbf{G} = (g_1, g_2, \cdots g_N), g_i \in \{-1, 1\}$ be the statistical phased haplotype. We can introduce statistical haplotype linkage into Eq. (6) as

$$E(\mathbf{S}) = \frac{1}{2} \sum_{1 \le i,j \le N} (M_{ij} + \rho_{ij} g_i g_j) s_i s_j. \quad [8]$$

Here ρ_{ij} is the "strength" of haplotype linkage from statistical phasing for sites i and j. Based on the decay of statistical phasing accuracy shown in Figure 1, we assume an exponential function of ρ_{ij} reflecting the accumulation of switching errors:

$$\rho_{ij} = e^{-d_{ij}/c}, \quad [9]$$

where d_{ij} is the 1D genomic distance between variant sites i and j and the constant c measures the average genetic distance of recombination. Because statistical phasing mainly relies on linkage disequilibrium between adjacent variants, we only incorporate linkage between adjacent variants. This leads to the following function to be maximized:

$$\begin{aligned} E(\mathbf{S}) &= \frac{1}{2} \left(\sum_{1 \le i,j \le N} M_{ij} s_i s_j + \sum_{1 \le i < N} \rho_i g_i g_{i+1} s_i s_{i+1} \right) \\ &= \frac{1}{2} \sum_{1 \le i,j \le N} M'_{ij} s_i s_j. \end{aligned} \quad [10]$$

Here $\rho_i = \rho_{i,i+1} = e^{-d_{i,i+1}/c}$.

C. Using Hi-C reads to correct long-range switching errors.

For haplotype determination using Hi-C and long reads, we previously maximized Eq. (6) by iteratively considering two perturbations to a random initial guess $\mathbf{S}_0$: a local spin flip

$$(s_1, s_2, \cdots, s_i, \cdots, s_N) \to (s_1, s_2, \cdots, -s_i, \cdots, s_N),$$

and a global switch

$$(s_1, s_2, \cdots, s_i, s_{i+1}, \cdots, s_N) \to (s_1, s_2, \cdots, s_i, -s_{i+1}, \cdots, -s_N).$$

To correct switching errors of large haplotype blocks, we consider "block flips" defined as

$$(s_1, s_2, \cdots s_k, s_{k+1} \cdots s_l, s_{l+1}, \cdots s_N) \to$$
$$(s_1, s_2, \cdots s_k, -s_{k+1} \cdots - s_l, s_{l+1}, \cdots s_N)$$

A block flip is equivalent to two global switches at site k and site l in a single step (Figure 1C). Starting with $\mathbf{S_0} = \mathbf{G}$, we find $\hat{\mathbf{S}}$ that maximizes Eq. (10) by applying block flips.

For a block flip between site $k+1$ and l, we have

$$\Delta E_{k|k+1,l|l+1} = -\sum_{k<i\le l} s_i \left(\sum_{j>l} M'_{mj} s_j + \sum_{j \le k} M'_{mj} s_j \right). \quad [11]$$

This can be calculated using a recursive relationship

$$\begin{aligned} \Delta E_{k|k+1,l|l+1} - \Delta E_{k-1|k,l-1|l} &= s_k \left(\sum_{i>l} + \sum_{i<k} - \sum_{k<i<l} \right) M'_{ik} s_i \\ &\quad - s_l \left(\sum_{i>l} + \sum_{i<k} - \sum_{k<i<l} \right) M'_{il} s_i. \end{aligned} \quad [12]$$

Note that $\Delta E_{0|1,l|l+1}$ is the block switching penalty $\Delta E_{l|l+1}$ given by

$$\Delta E_{l|l+1} = -\sum_{i \le l} \sum_{j>l} M'_{ij} s_i s_j. \quad [13]$$

and can be calculated using

$$\Delta E_{l|l+1} - \Delta E_{l-1|l} = s_l \left(\sum_{j>l} M'_{lj} s_j - \sum_{i<l} M'_{il} s_i \right). \quad [14]$$

We use this recursion to calculate block flipping penalties. For each step of haplotype refinement, block flipping penalties are calculated at every variant position for block sizes ranging from half the total number of variants on a chromosome $(N/2)$ down to a single variant. Specifically, we calculate large-scale flipping penalties for block sizes in the sequence $N/2, N/4, \ldots, \lceil N/2^{i+1} \rceil > 2000$.

Table 1. Genome-wide phasing accuracy

Genome	phased in truth set	linked by Hi-C	comparable	accuracy	EAGLE2 recovery	comparable	accuracy
NA12878	2,156,346	1,650,419	1,611,090	0.997	2,043,936	1,784,694	0.997
RPE-1	2,151,642	1,819,693	1,736,800	0.992	2,484,304	2,061,065	0.991
HCC1954BL	2,027,293	2,156,067	1,863,261	0.993	2,459,343	1,881,245	0.992
HCC1954	2,027,293	2,010,935	1,739,987	0.994	2,459,230	1,881,245	0.994

Finer-scale flipping penalties are calculated for block sizes on the interval $[2000, 200]$ with a 20 variant decrement and on the interval $[200, 1]$ with a 1 variant decrement.

This results in a grid of block flipping penalties for every position on the chromosome, for every block size described above. We then simultaneously flip the haplotype assignment for disjoint blocks with the most favorable switching penalties satisfying $\Delta E \geq 10.0$. This process is repeated until no remaining favorable block flips are found.

The resulting solution only includes phased genotypes at variant sites with Hi-C linkage (see Table 1,"linked by Hi-C"). For variants not having Hi-C links, we determined their haplotype phase based on their statistical phasing linkage to the nearest Hi-C phased variant (up to 5kb, see Table 1 "EAGLE2 recovery"). The combination of Hi-C phasing and local haplotype recovery produces a chromosomal haplotype with phased genotypes at $> 99\%$ common variant sites (~ 1 variant per 1.5 kb).

D. Benchmarking haplotype inference with refLinker. We ran `refLinker` on three diploid human genomes with ground truth haplotype information to benchmark the accuracy and completeness of haplotype inference from Hi-C data (see Table S1 for Hi-C data sources). For NA12878, truth haplotypes were determined using parental genomes in the Genome In A Bottle release (20); for RPE-1 cells, truth haplotypes were determined from single-cell sequencing and independently validated by computational inference from linked-reads and Hi-C with `mLinker` (11); for HCC1954BL, truth haplotypes were determined from linked-reads and Hi-C with `mLinker` and validated by allelic imbalance in the matching breast cancer genome HCC1954.

Phasing accuracy was evaluated both for variants with direct haplotype linkage from Hi-C reads (Table 1: "linked by Hi-C") and including variants with only statistical haplotype linkage to Hi-C phased variants (Table 1: "recovered"). In all cases, `refLinker` was able to resolve the haplotype phase of common germline variants at $> 99\%$ accuracy (Table S2, S3, S4). To determine the minimum number of Hi-C reads required for long-range phasing accuracy, we ran `refLinker` on downsampled Hi-C data of RPE-1 cells ($0.333\times$ and $0.167\times$; starting library: $245,700,771$ reads, 29.5 million long-range contacts). `refLinker` achieved $> 98\%$ phasing accuracy up to $0.333\times$ downsampling (Table S5), but we noted decreasing accuracy for some chromosomes at $0.167\times$ downsampling (Table S6). We conclude that `refLinker` requires ~ 10 million long-range Hi-C contacts (between loci separated by ≥ 1Mb) for robust, long-range haplotype inference.

Next, we compared `refLinker`'s phasing accuracy to IntegratedPhasing (16). IntegratedPhasing and `refLinker` differ in both (1) their approaches for integrating haplotype linkage from statistical phasing; and (2) their quantification of Hi-C linkage. For diploid phasing applications, the first point is most relevant: because IntegratedPhasing uses the SHAPEIT2 haplotype graph to extract high-confidence statistical phasing linkage, it is less flexible to the incorporation of improved statistical phasing algorithms, such as EAGLE2 (Figure 1A). For comparison, we ran IntegratedPhasing on Hi-C data from the RPE-1 genome (Table S7). After running on complete Hi-C data from RPE-1, IntegratedPhasing achieved 98.7% accuracy. However, IntegratedPhasing's lower overall accuracy on downsampled Hi-C data (95.9%, including seven chromosomes with $> 5\%$ phasing errors after $0.333\times$ downsampling; Table S7) indicates that IntegratedPhasing requires more long-range Hi-C contacts than `refLinker` for robust, long-range haplotype inference.

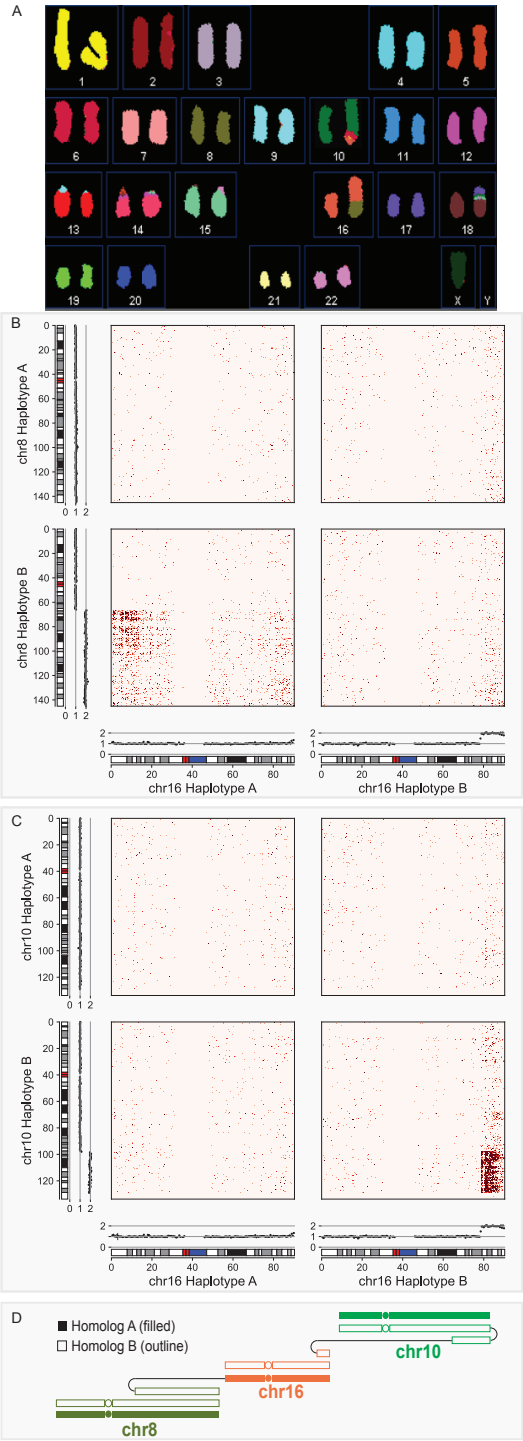

Fig. 2. `refLinker` resolves the parental origin and composition of translocations in HCT116. (A) SKY karyotype of HCT116 from the Pawefish project used with permission from Dr. Paul Edwards https://www.cellosaurus.org/pawefish/ColonCellLineDescriptions/HCT116.html. (B) Haplotype-specific DNA copy number and inter-chromosomal Hi-C contacts of chr8 and chr16. The homologs of chr8 are arranged vertically and chr16 homologs are arranged horizontally. DNA copy number of each homolog is shown as a black dots. Hi-C contacts are shown as heatmaps. (C) Haplotype-specific copy number and interchromosomal Hi-C contacts between chr10 (vertical axis) and chr16 (horizontal axis). Data are arranged similarly to (B). (D) Schematic drawing of two derivative chromosomes involving different chr16 homologs. Homolog A is indicated by filled shapes and homolog B is indicated by outlines of the same color.

Finally, we tested each method's robustness across large variant deserts or gaps, including centromeres. Having established that `refLinker` requires ~ 10 million long-range Hi-C contacts for robust haplotype inference, we ran both `refLinker` and IntegratedPhasing on $0.333\times$ downsampled Hi-C data from RPE-1. We then examined switching errors across HSat3 on chr9, a 27.6 Mb repetitive array that leads to spurious errors between p- and q-arm haplotypes (14). For 20 independent trials of $0.333\times$ Hi-C downsampling, three runs from IntegratedPhasing returned arm-level switching errors (Figure S1A, left), whereas `refLinker` returned none (Figure S1A, right). The accuracy of `refLinker` haplotypes ranged from $97.5 - 98.6\%$ (Figure S1B).

In summary, the NA12878, RPE-1, and HCC1954BL benchmarking studies demonstrated the consistent accuracy of `refLinker` in chromosomal haplotype inference, with $> 99\%$ phasing accuracy for $> 99.9\%$ of common variants from ~ 250 million Hi-C reads (≥ 10 million long-range Hi-C contacts). In comparison to existing tools, `reflinker` shows the most robust long-range phasing accuracy, with comparable completeness in the final haplotype.

E. Determination of preexisting and de novo rearrangements in HCT116. To demonstrate `refLinker`'s utility for cancer genome reconstruction, we considered the widely studied HCT116 genome. HCT116 cells are near-diploid with three translocations visible by standard karyotyping: der(10)dup(10)(q?)t(10;16), der(16)t(8;16), and der(18)t(17;18) (Figure 2A) (21). Despite its widespread use in genomic studies, haplotype phasing of HCT116 has never been attempted. Noting the availability of whole-genome and Hi-C sequencing for HCT116 (22, 23), we set out to reconstruct the HCT116 genome using `refLinker`.

We ran EAGLE2 on common polymorphisms in HCT116 and ran `refLinker` using Hi-C data from a previous study (23). The final haplotype phase contained $1,757,238$ phased variants (Dataset S1). The number of phased variants for HCT116 is lower than the previous three genomes used for benchmarking due to large regions of copy-number neutral loss-of-heterozygosity (LOH) on the p-arms of chr3, chr5, and chr7 (Figure S2). We note that the parental origin of the retained haplotype in these LOH regions cannot be established as the same genotypes are present on both homologs for these three cases.

We then calculated the DNA copy-number of every parental chromosome, revealing haplotype-specific duplications on chromosome 8, 10, 16, and 17 (Figure S3). Based on the haplotype phase, we further extracted haplotype-specific Hi-C contacts to resolve the structure of translocated chromosomes in HCT116 (Figure 2B-D and Figure 3A).

The first translocation occurs between a duplicated q-terminal segment of the 8B homolog and the p-terminus of 16A. This translocation is labeled 16 in the SKY image (Figure 2A) and the structure is supported by the enrichment of Hi-C contacts between the 16p terminus and 63 Mb on 8q (Figure 2B). The second translocation occurs between a duplicated q-terminal segment of the 16B homolog and the q-terminus of the 10B homolog (Figure 2C). In this case, there is a pre-existing terminal duplication on the 10q arm that is inferred to be an inverted duplication (or foldback). The structure of the translocated chromosome is inferred from increased Hi-C contacts at the 10q-terminus (Figure S4A) and DNA sequencing reads supporting the chr10:98,953,544-chr16:79,696,003 junction (Figure S4B). The phasing of two translocations to different chr16 homologs suggests that these events occurred independently. Finally, a third translocation occurs

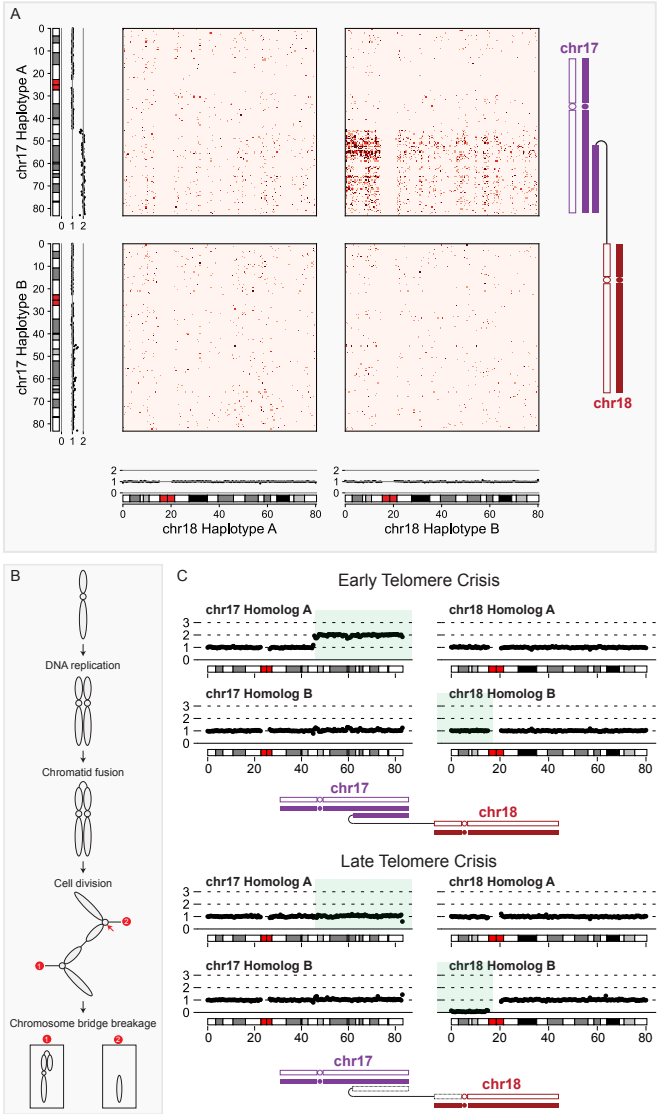

Fig. 3. Digital karyotyping resolves de novo copy-number alterations in HCT116 cells following telomere crisis. (A) Phased Hi-C contacts reveal a chr17A-chr18B translocation. Data are shown similarly as in Figure 2 for chr17 (vertical) and chr16 (horizontal) Right: structure of the derivative chromosome. (B) Schematic of reciprocal arm-level gain and loss resulting from a single round of chromatid-type breakage-fusion-bridge cycle. (C) chr17A-chr18B linkage revealed in HCT116 evolution following telomere crisis. Homolog specific copy-number was calculated from whole-genome sequencing of HCT116 clones at two timepoints after escape from telomere crisis. Top: Whole-genome sequencing of SRR18252999, HCT116 early telomere crisis clone, reveals chr17 and chr18 copy-number consistent with HCT116 karyotype. Drawn schematics indicate the two homologs of chr17 and molecular linkage between the chr17A duplicated segment with chr18B. Bottom: Whole-genome sequencing of the same clone at a later timepoint (SRR18252998) reveals a single-copy loss of the chr17A duplicated segment (top plot) linked to loss of the chr18B q-arm (bottom plot). Drawn schematics indicating arm-level loss on the chr17A-chr18B derivative chromosome consistent with HCT116 haplotype reconstruction by refLinker. Lost segments are indicated by gray boxes.

between a duplicated q-terminal segment of 17A and the 18B homolog (Figure 3A). These analyses completely determine the haplotype identity and breakpoints of all three translocations.

We next used the haplotype phase and haplotype-resolved karyotype of HCT116 cells to analyze *de novo* alterations after telomere crisis using standard shotgun sequencing data (22, 24). We calculated haplotype-specific DNA copy number from whole-genome sequencing data generated on early- and late-passage clones following telomere crisis. A common alteration after telomere crisis is arm-level deletion and an example is shown in Figure 3B and C. In this example, the late-passage clone showed concurrent loss of the 18B p-arm and the 17A q-terminal segment that was inferred to be translocated to the 18B p-terminus. Based on the haplotype-resolved karyotype, we concluded that both deletions occurred as a single arm-level deletion.

We further identified an example of copy-number evolution indicating successive breakage-fusion-bridge (BFB) cycles on the 11B homolog (Figure S5). In this example, an ancestral internal deletion on 11B as identified in an early passaging clone was inferred to be followed by secondary duplications and deletions at the centromeric break end. The pattern is consistent with a chromosome-type BFB cycle (25) followed by secondary fusion/breakage events.

F. Haplotype resolution of complex rearrangements in the HCC1954 genome. We then studied the highly rearranged HCC1954 breast cancer cell line to evaluate `refLinker`'s utility for complex genome reconstruction. In comparison to HCT116, HCC1954 is near-tetraploid and contains extensive copy-number alterations and structural rearrangements that disrupt parental homolog structure, posing challenges for accurate haplotype inference. HCC1954 is well-suited for this benchmarking task because ground-truth haplotypes have been determined for its matching germline reference, the diploid HCC1954BL cell line (Table S4). Evaluating refLinker's haplotype inference in HCC1954 against the HCC1954BL ground-truth reference, we found that `refLinker` phases germline variants in HCC1954 with 99.4% accuracy (Table S8).

We then determined the chromosomal locations of rearranged DNA from chr17, including highly amplified copies (> 100) of the *ERBB2* gene. Using the parental haplotype generated from `refLinker`, we calculated the copy-number of both chr17 homologs and identified segmental copy-number alterations confined to a single chr17 homolog (17A). These include complete deletion of the p-arm, which results in monoallelic expression of mutant p53, large segmental deletions on the q-arm, and high-level amplification of the *ERBB2* oncogene (Figure 4A, B, top). The second chr17 homolog (17B) is present in four structurally intact copies (Figure 4A, B, bottom).

Based on haplotype-resolved Hi-C contacts, we determined the locations of 17q material in three marker chromosomes (Figure 4B-E). Large retained segments on 17q are translocated to chr8 27-29 Mb, 3 copies, $\sim 62 - 72.5$ Mb, 3 copies, and $\sim 82 - 83.5$ Mb, 3 copies; Figure 4C) and chr13 $\sim 29 - 43$ Mb 1 copy and ~ 72.5–81.5 Mb, 1 copy; Figure 4D). Separately, the *ERBB2* amplicon is located in a homogeneously staining region (HSR) that is translocated to chr12 (Figure 4E). These findings are validated by spectral karyotyping (SKY). Notably, SKY cannot resolve the DNA content of the *ERBB2* HSR due to the repetitive structure of the amplified DNA and the small size of the 12p-terminal segment. The content of the HSR chromosome is validated by a rearrangement junction between 12p and the amplified region in 17q together with

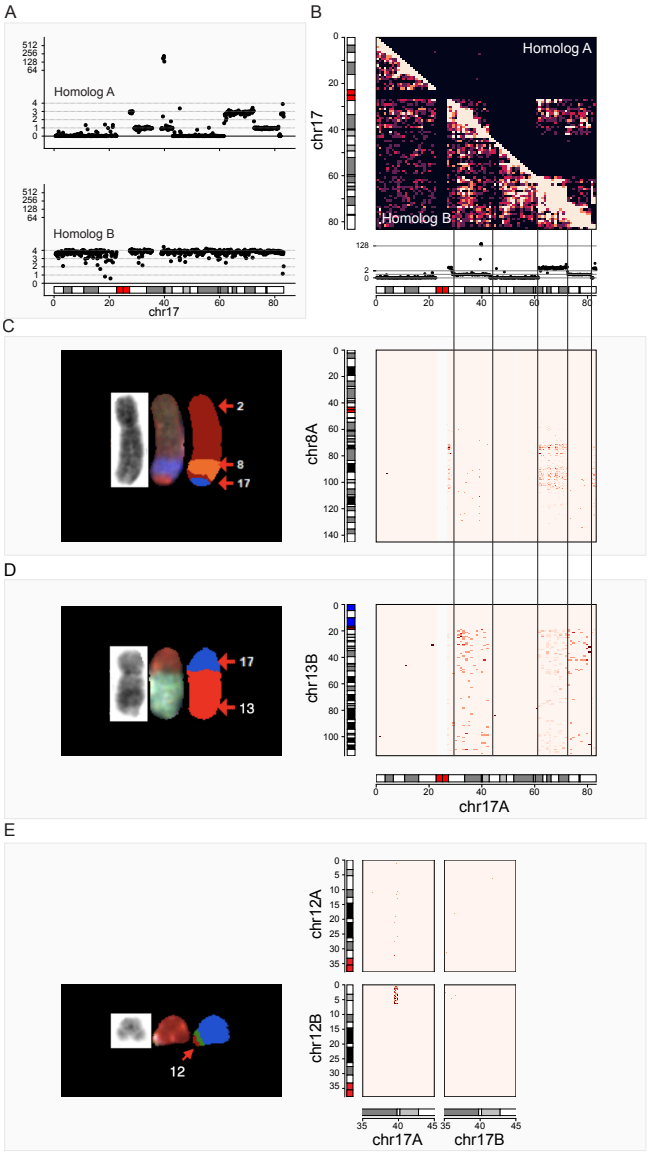

Fig. 4. Haplotype-resolved genome complexity in HCC1954. (A) Allele-specific copy-number of HCC1954 chr17. (B) Top: Hi-C maps of HCC1954 chr17 intrachromosomal interactions for homolog A (upper triangle) and homolog B (lower triangle). Bottom: DNA copy number of HCC1954 chr17A, revealing large deletions and high-copy *ERBB2* amplification. (C) Hi-C maps between both chr8 homologs (vertical axis) and chr17 homologs (horizontal axis), revealing chr17A translocation to chr8A. (D) Hi-C maps between both chr13 homologs (vertical axis) and chr17 homologs (horizontal axis) revealing chr17A translocation to chr13B. (E) Hi-C maps placing the *ERBB2* amplicon in a derivative chromosome capped by chr12B telomere. For C, D, and E, Hi-C linkage is normalized by chr17 allelic copy-number.

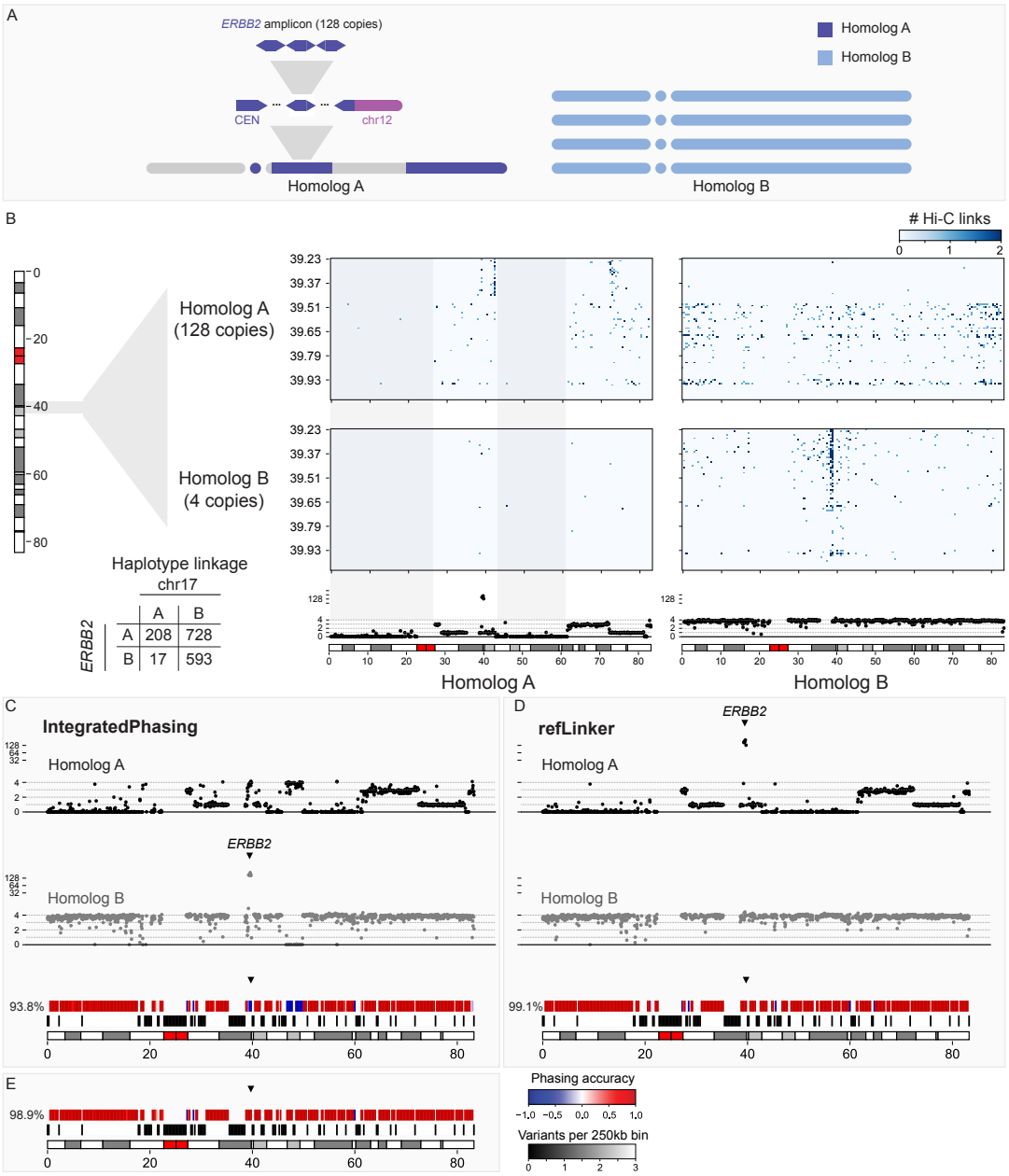

Fig. 5. (A) Schematic showing HCC1954 chr17 A and B homologs (dark blue and light blue, respectively). Haplotype resolved DNA copy-number reveals Homolog A is highly rearranged with p-arm deletion, large deletions on the q-arm and and *ERBB2* focal amplification (left). In contrast, Homolog A is present in four intact copies (right). (B) Haplotype-specific Hi-C contact map between the *ERBB2* locus and the rest of chr17, where the total number of Hi-C contacts are: 17B-*ERBB2*↔17A (728), 17A-*ERBB2*↔17A (593), 17B-*ERBB2*↔17B (208) and 17A-*ERBB2*↔17B (17). (C) IntegratedPhasing was run on HCC1954 Hi-C data. Allelic copy-number is plotted for chr17 Homolog A (top row, gray) and Homolog B (second row, black). Phasing accuracy and variant density is plotted for 250 kb bins (bottom two tracks). Whole-chromosome IntegratedPhasing accuracy for chr17 was 93.8%. Focal amplification containing *ERBB2* was assigned to Homolog A (black arrow). (D) refLinker was run on HCC1954 Hi-C data and initialized with EAGLE2 haplotypes for HCC1954BL. Allelic copy-number is plotted for chr17 Homolog A (top row, gray) and Homolog B (second row, black). Phasing accuracy and variant density is plotted for 250 kb bins (bottom two tracks). Whole-chromosome refLinker phasing accuracy for chr17 was 99.1%. Focal amplification containing *ERBB2* was assigned to Homolog B (black arrow). (E) refLinker was run on HCC1954 Hi-C data and initialized with IntegratedPhasing haplotype from (G). Phasing accuracy (top track)and variant density (bottom track) is plotted for 250 kb bins. Whole-chromosome refLinker phasing accuracy for chr17 was 98.9%. Phasing accuracy at the *ERBB2* locus is indicated by the black arrow.

the size of the 12p-terminal segment in the marker chromosome determined from the copy number (Figure 4A, E).

Because these deletions and translocations destroy the linear structure of chr17B, the *ERBB2* amplicon (hereafter called *ERBB2*-HSR) has completely lost all *cis* linkage to the original chr17 homolog. Additionally, the high copy-number of *ERBB2*-HSR causes abundant non-specific Hi-C contacts with the opposite, intact homolog. This is illustrated by the haplotype-specific Hi-C contact map between the *ERBB2* locus and the rest of chr17 (Figure 5B). In total, the nonspecific *trans* contacts between *ERBB2*-HSR and the opposite homolog (728 contacts) exceeds the number of proximity-based cis contacts on either haplotype (208 contacts for Homolog A, 593 contacts for Homolog B) (Figure 5B, left).

The phasing of HCC1954 chr17 highlights the unique challenges of haplotype phasing in complex genomes and demonstrates the performance differences between `refLinker` and IntegratedPhasing. As IntegratedPhasing considers the four classes of haplotype linkage (AA, AB, BA, and BB) separately and switches haplotype assignments based on the dominant linkage, it incorrectly assigns the 17A-*ERBB2* allele to the 17B haplotype (Figure 5C). By contrast, our formalism considers combined linkage AA + BB versus AB + BA, which leads to the correct haplotype assignment as 208 (AA) $+$ 593 (BB) $=$ 801 $>$ 728 (AB) $+$ 17 (BA) (Figure 5D, E). Additionally, beyond considering the raw magnitude of haplotype linkage, the introduction of the error-correction prefactor in Eq. 5 can further penalize trans linkage (728 (AB) and 17 (BA)). Therefore, our formalism enables a more robust statistical framework for accruing haplotype linkage evidence in complex genomes.

1. Discussion

Here, we present `refLinker`, a computational method that enables chromosome-scale haplotype inference in complex genomes. We evaluate `refLinker`'s performance in five human genomes, including two aneuploid cancer genomes. In contrast to prior methods, `refLinker` requires only Hi-C data and provides whole-chromosome haplotype inference with comparable or superior completeness and accuracy to methods requiring multiple long reads and Hi-C sequencing modalities (11, 15). The consistency of long-range haplotype inference by `refLinker` is also demonstrated for highly aneuploid and structurally complex genomes. This capability further enables the resolution of the long-range structure of rearranged chromosomes by haplotype-specific Hi-C contacts. The combination of long-range haplotype phasing and the resolution of long-range chromosomal structure demonstrates a unique application of Hi-C sequencing for long-range cancer genome analysis, which we anticipate will have direct applications in relating changes in DNA sequence, epigenetic state, and gene expression to large-scale chromosomal rearrangements.

`refLinker` provides several advantages over IntegratedPhasing, the only comparable method for haplotype phasing with this datatype. First, `refLinker` combines the Hi-C linkage from both parental homologs to reach a consensus haplotype assignment. This provides more stable solutions than IntegratedPhasing, which separately calculates molecular linkage for each homolog to assign haplotype phase. `refLinker`'s integrated approach achieves $> 99\%$ phasing accuracy even when only one parental chromosome is retained intact (see Section F). Second, `refLinker` can be run with the phased variant output of any statistical phasing method, allowing straightforward integration of future improvements in statistical phasing algorithms. As the size of genotyped cohorts grows sufficiently large, the increasing accuracy of statistical phasing may eventually eliminate the need for reads-based phasing approaches for common variants.

Our approach for Hi-C-based error correction may be applied more broadly to the problem of *de novo* assembly in cancer genomes, where large, duplicated segments are often assigned to the wrong contigs. We note that `refLinker`'s approach provides an automatic procedure for the identification and correction of these errors using Hi-C. Just as with errors in haplotype assignment, assembly errors create a localized enrichment of spurious inter-chromosomal Hi-C contacts. By identifying these genomic segments and switching their contig assignment to minimize the total number of inter-chromosomal interactions, `refLinker`'s approach provides a framework for systematic improvement of *de novo* assemblies. Together with our results, these applications illustrate `refLinker`'s broad utility and the applications of its underlying approach to future studies.

Acknowledgments

We thank Dr. Paul Edwards for generously allowing us to re-print the HCT116 SKY karyotype. G.J.B. was supported by the training grant T15LM007092 from the National Library of Medicine. C.-Z. Z. was supported by a NCI K22 award (K22CA216319) and by the Claudia Adams Barr Program for Innovative Cancer Research from the Dana-Farber Cancer Institute. D.P. is a HHMI Investigator, a member of the HMS/Boston Ludwig Center and is supported by NIH R01CA213404-24 and an award from the G. Harold and Leila Y. Mathers Foundation.

1 JD Rowley, A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature* **243**, 290–293 (1973).
2 DF Robbiani, MC Nussenzweig, Chromosome translocation, b cell lymphoma, and activation-induced cytidine deaminase. *Annu. Rev. Pathol. Mech. Dis.* **8**, 79–103 (2013).
3 CR King, MH Kraus, SA Aaronson, Amplification of a novel v-erb b-related gene in a human mammary carcinoma. *Science* **229**, 974–976 (1985).
4 S Nurk, et al., The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
5 H Cheng, GT Concepcion, X Feng, H Zhang, H Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. methods* **18**, 170–175 (2021).
6 M Rautiainen, et al., Telomere-to-telomere assembly of diploid chromosomes with verkko. *Nat. Biotechnol.* pp. 1–9 (2023).
7 WW Liao, et al., A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
8 H Cheng, et al., Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).
9 A Adey, et al., The haplotype-resolved genome and epigenome of the aneuploid hela cancer cell line. *Nature* **500**, 207–211 (2013).
10 Y Ghavi-Helm, et al., Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. genetics* **51**, 1272–1282 (2019).
11 RW Tourdot, GJ Brunette, RA Pinto, CZ Zhang, Determination of complete chromosomal haplotypes by bulk DNA sequencing. *Genome Biol.* **22**, 139 (2021).
12 R Schöpflin, et al., Integration of hi-c with short and long-read genome sequencing reveals the structure of germline rearranged genomes. *Nat. Commun.* **13**, 6470 (2022).
13 J Ijaz, et al., Haplotype-specific assembly of shattered chromosomes in oesophageal adenocarcinomas. *Cell Genomics* (2024).
14 S Selvaraj, JR Dixon, V Bansal, B Ren, Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. biotechnology* **31**, 1111–1118 (2013).
15 P Edge, V Bafna, V Bansal, Hapcut2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome research* **27**, 801–812 (2017).
16 V Bansal, Integrating read-based and population-based phasing for dense and accurate haplotyping of individual genomes. *Bioinformatics* **35**, i242–i248 (2019).
17 O Delaneau, JF Zagury, J Marchini, Improved whole-chromosome phasing for disease and population genetic studies. *Nat. methods* **10**, 5–6 (2013).
18 PR Loh, et al., Reference-based phasing using the haplotype reference consortium panel. *Nat. genetics* **48**, 1443 (2016).

19 GPCC authors Auton Adam adam. auton@ gmail. com 1 b Abecasis Gonçalo R. goncalo@ umich. edu 2 c, et al., A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
20 S Fairley, E Lowy-Gallego, E Perry, P Flicek, The international genome sample resource (igsr) collection of open human genomic variation resources. *Nucleic acids research* **48**, D941–D947 (2020).
21 WM Abdel-Rahman, et al., Spectral karyotyping suggests additional subsets of colorectal cancers characterized by pattern of chromosome rearrangement. *Proc. Natl. Acad. Sci.* **98**, 2538–2543 (2001).
22 K Cleal, RE Jones, JW Grimstead, EA Hendrickson, DM Baird, Chromothripsis during telomere crisis is independent of nhej, and consistent with a replicative origin. *Genome research* **29**, 737–749 (2019).
23 S Rao, et al., Cohesin loss eliminates all loop domains. *Cell* **171**, 305–320 (2017).
24 K Liddiard, AN Aston-Evans, K Cleal, EA Hendrickson, DM Baird, Polq suppresses genome instability and alterations in dna repeat tract lengths. *NAR cancer* **4**, zcac020 (2022).
25 NT Umbreit, et al., Mechanisms generating cancer genome complexity from a single cell division error. *Science* **368**, eaba0712 (2020).