GAMIBHEAR: whole-genome haplotype reconstruction from Genome Architecture Mapping data

Julia Markowski[1,2], Rieke Kempfer[1,2], Alexander Kukalev[1], Ibai Irastorza-Azcarate[1] , Gesa Loof[1,2], Ana Pombo[1,2], Roland F Schwarz[1,#]

[1] Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin
[2] Department of Biology, Humboldt University of Berlin, Germany

[#] corresponding author: roland.schwarz@mdc-berlin.de

Abstract

Motivation

Understanding haplotype-specific regulatory mechanisms becomes increasingly important in genomics and medical research. Investigating differences in allele-specific gene expression, epigenetic changes and their causal variants greatly benefits from haplotype reconstruction or phasing of genetic variants, but direct evidence for the haplotype structure is difficult to obtain from standard short-read sequencing data. Chromatin conformation data obtained from 3C experiments allows inference of haplotypes because inter-chromosomal contacts are more frequent than homologous intra-chromosomal contacts, but these data suffer from technical biases owing to the digestion and ligation process of the 3C technique. Genome Architecture Mapping (GAM) is a novel digestion- and ligation-free method for the inference of chromatin conformation from nuclear cryosections. Due to its high resolution and independence of enzymatic digestion it is well-suited for haplotype reconstruction and for detecting haplotype-specific chromatin contacts.

Results

Here, we present GAMIBHEAR, a tool for accurate haplotype reconstruction from GAM data. GAMIBHEAR aggregates allelic co-observation frequencies across multiple nuclear slices and employs a GAM-specific probabilistic model of haplotype capture to optimise phasing accuracy. Using a hybrid mouse embryonic stem cell line with known haplotype structure as a benchmark dataset, we assess correctness and completeness of the reconstructed haplotypes, and demonstrate the power of GAM data and the accuracy of GAMIBHEAR to infer genome-wide haplotypes.

Availability

GAMIBHEAR is available as an R package under the open source GPL-2 license at https://bitbucket.org/schwarzlab/gamibhear
Maintainer: julia.markowski@mdc-berlin.de

Introduction

Assigning genetic variants identified by short-read sequencing to their physical parental chromosomal copies or haplotypes is known as phasing and is a key challenge in genomics research. Haplotypes are invaluable in many application areas, such as genotype imputation (Tewhey et al. 2011), in identifying additive effects of genetic variants to the regulation of gene expression in human disease (PCAWG Transcriptome Core Group et al. 2020) and in detecting somatic chromosomal aberrations in cancer (Jamal-Hanjani and Wilson 2017). Traditionally, haplotypes are inferred through read-based phasing methods (Bansal and Bafna 2008; Patterson et al. 2015) or statistically using population-level or reference-phasing approaches (Loh et al. 2016; Browning and Browning 2007).

Recently, attempts have been made to leverage chromatin conformation data for haplotype reconstruction(Chaisson et al. 2019). Chromatin is spatially well organized in the nucleus in a hierarchical and structured way. Most experimental techniques that reveal chromatin structure, such as the popular 3C-derivative Hi-C (Lieberman-Aiden et al. 2009), are based on crosslinking DNA followed by restriction-enzyme initiated digestion and proximity-based re-ligation of the DNA fragments. Ligation products of interacting genomic regions are identified through next-generation sequencing and give rise to chromatin contact maps. Motivated by earlier observations that homologous chromosomes tend to occupy distant chromosome territories (Meaburn and Misteli 2007), Selvaraj et al. (2013) demonstrated that homologous interchromosomal (*h-trans*) contacts are exceedingly rare compared to intrachromosomal (*cis*) contacts. To leverage this phenomenon, the authors proposed HaploSeq, a combination of the Hi-C protocol with computational haplotype assembly using the HapCut algorithm (Bansal and Bafna 2008).

However, technical biases in 3C-type methods introduced by the digestion and ligation process, such as non-uniformity in sequence coverage (Bansal 2019) and the technical limitation to two- or three-way contacts (O'Sullivan et al. 2013) impair the accuracy and completeness of reconstructed haplotypes. For example, employing HaploSeq, Selvaraj *et al.* (2013) reported overall low resolution from Hi-C reads alone, with about 22% of variants phased in a human genome. The authors, and later Bansal *et al.* (2019), attempted to overcome this by incorporating statistical phasing into the reconstruction algorithm, drastically increasing the resolution. Despite these successes, chromatin contact data unaffected by ligation and digestion biases would be highly desirable for the inference of haplotypes and contact maps alike.

Genome Architecture Mapping (GAM) is a novel digestion- and ligation-free experimental technique for assessing the 3D chromatin structure from a collection of thin nuclear profiles (NPs) (Beagrie et al. 2017). NPs are generated through cryosectioning of cellular nuclei followed by next-generation sequencing. Chromatin contacts between DNA loci can then be inferred by analyzing their co-observation frequency, i.e. the frequency at which the loci are captured in the same NP. In contrast to 3C-type approaches, GAM is able to resolve complex contacts with three or more loci with high resolution, does not suffer from non-uniformity biases and only requires several hundreds of cells to obtain high-resolution contact maps (Kempfer and Pombo 2019).

These advantages should make GAM particularly useful for haplotype reconstruction and allele-specific analyses of chromatin contacts in rare biological materials, such as human biopsies. Unfortunately, no study has so far investigated the use of sequencing data from nuclear sections

for variant phasing and haplotype reconstruction and no algorithm exists for inferring haplotype structure from GAM data.

To address this, we present GAMIBHEAR (GAM-Incidence Based Haplotype Estimation And Reconstruction), a novel computational tool for whole-genome phasing of genetic variants from GAM NPs. GAMIBHEAR employs a graph representation of the co-occurence of SNP alleles in the nuclear profile to reconstruct the haplotype structure. It thereby accounts for the GAM-specific probabilities in capturing parental chromosomal segments as part of the random cryosectioning process. We assess the performance of GAMIBHEAR on the hybrid mouse embryonic stem cell line (clone F123) with known haplotype structure and demonstrate its ability to reconstruct accurate and complete whole-genome haplotypes. GAMIBHEAR is available as an efficient R package with parallel implementations of the most compute-intensive tasks and is available on https://bitbucket.org/schwarzlab/gamibhear.

Materials and methods

Dataset

We use the hybrid mouse embryonic stem cell line (clone F123) as a benchmark system for assessing the quality of reconstructed haplotypes from GAM data. The F123 line is derived from the F1 generation of two fully inbred homozygous mouse strains: *Mus musculus castaneus* (CAST) and 129S4/SvJae (J129) (Gribnau et al. 2003). Whole-genome sequencing (WGS) data of CAST and J129 were downloaded from the European Nucleotide Archive (accession number ERP000042) and the Sequence Read Archive (accession number SRX037820). To determine the haplotypes of the F123 line, WGS reads were trimmed using Cutadapt (Martin 2011) and mapped to mm10 using BWA (Heng Li and Durbin 2009). SNPs were identified using bcftools (Heng Li 2011) and SNPs covered by <5 reads and quality <30 were excluded.

Relative to the mouse reference genome mm10, CAST and J129 show 18,892,144 and 4,778,766 germline variants respectively, in concordance with their estimated evolutionary distance from mm10 (371,000 ± 91,000 years (Goios et al. 2007) and approximately 100 years (Simpson et al. 1997), respectively). After exclusion of 2,200,819 overlapping SNP positions, the F123 reference set contains 19,269,272 variants in total, all of which are heterozygous due to inbreeding of the parental strains. This yields an average SNP density of 1 SNP per 125bp. With the haplotype structure thus known, this cell line serves as the benchmark for all downstream experiments and analyses.

GAM pre-processing and quality control

1261 GAM samples from individual nuclear profiles (NPs) of the F123 line were obtained from the 4D Nucleome Consortium data portal under accession number 4DNBSTO156AZ. F123 SNPs were N-masked in the mm10 reference genome and reads were mapped using Bowtie2 (Langmead and Salzberg 2012). Duplicate reads were removed using samtools (H. Li et al. 2009). After mapping, all BAM files and WGS results underwent standard quality control using FastQC (Andrews 2010) and multiQC (Ewels et al. 2016). Reads were trimmed using BamUtil (Jun et al. 2015) with function trimBam where necessary.

GAM data from single NPs covers a proportion of the whole genome with consecutive stretches of genomic DNA that reflect chromatin looping in and out of a thin nuclear slice. For quality

assessment of each sample, the genome was split into fixed windows of size 50kb. For each NP i and each window j, the number of reads r_{ij} and number of nucleotides covered c_{ij} were determined using bedtools (Quinlan and Hall, 2010). Windows were then classified as *positive* or *negative* based on r_{ij} and c_{ij} as follows: from the coverage $c_{i\bullet}$ of all windows for NP i the empirical nucleotide coverage distribution P_i was computed. From P_i the minimum coverage percentile MCP_i was chosen such that every window contains three or more reads. The average $\overline{MCP}$ across all NPs then determined the sample-specific nucleotide coverage thresholds t_i (in bp) for each NP. Windows w_{ij} were called positive iff $c_{ij} > t_i$, i.e. if the number of nucleotides covered in each window was greater than the sample-specific threshold and negative otherwise. *Positive* windows flanked by *negative* windows on each side were defined as *orphan* windows. NPs selected for further analysis had < 60% orphan windows and > 20,000 uniquely mapped reads. 1123 NPs passed the quality thresholds.

Reads were then counted at known heterozygous SNP positions using samtools mpileup (H. Li et al. 2009).

Probabilistic model of the GAM sectioning process

Each NP in GAM is the result of random sectioning of the nucleus and captures ultra-sparse *local* sequence information, where *local* refers to genomic loci in close proximity in the 3D arrangement of the genome, including loci proximal in linear distance. We extracted on average 305,377 reads from each NP, covering 0.166 % (std. error = 0.0048) of the 19,269,272 heterozygous SNPs per nuclear slice (Figure 1A). Out of these SNPs, 12,155,703 (63.08 %) were observed at least once across all 1123 NPs and 7,813,796 SNPs (40.55 %) were observed at least twice (Figure 1B). Due to this sparsity and the fact that homologous chromosome pairs occupy distinct chromosomal territories (Khalil et al. 2007), 96.57% of SNP observations showed counts from only one parental allele. Thus,we removed observed variants with read counts from both parental alleles without substantial loss of information.

Indeed, alleles of any two SNPs captured in a nuclear slice were more likely to originate from the same parental copy and this likelihood decreases with increasing genomic distance of the co-observed alleles (Figure 1C). To account for the decay of this local phasing signal with increasing genomic distance, we modelled the empirical probability p of two alleles coming from the same haplotype based on their genomic distance d using least squares and the exponential model $p \sim a \cdot e^{(b \cdot d)}$. For this model we only considered SNPs with distance $100\,kb < d < 10\,Mb$, where the decay in phasing information is most pronounced (Figure 1D). For variants outside that range, probabilities 1 and 0 were assumed respectively. Parameter $a = 0.974$ then describes the co-observation probability at a genomic distance of 100 kb with an exponential decay of $b = -3.173 \cdot 10^{-7}$.

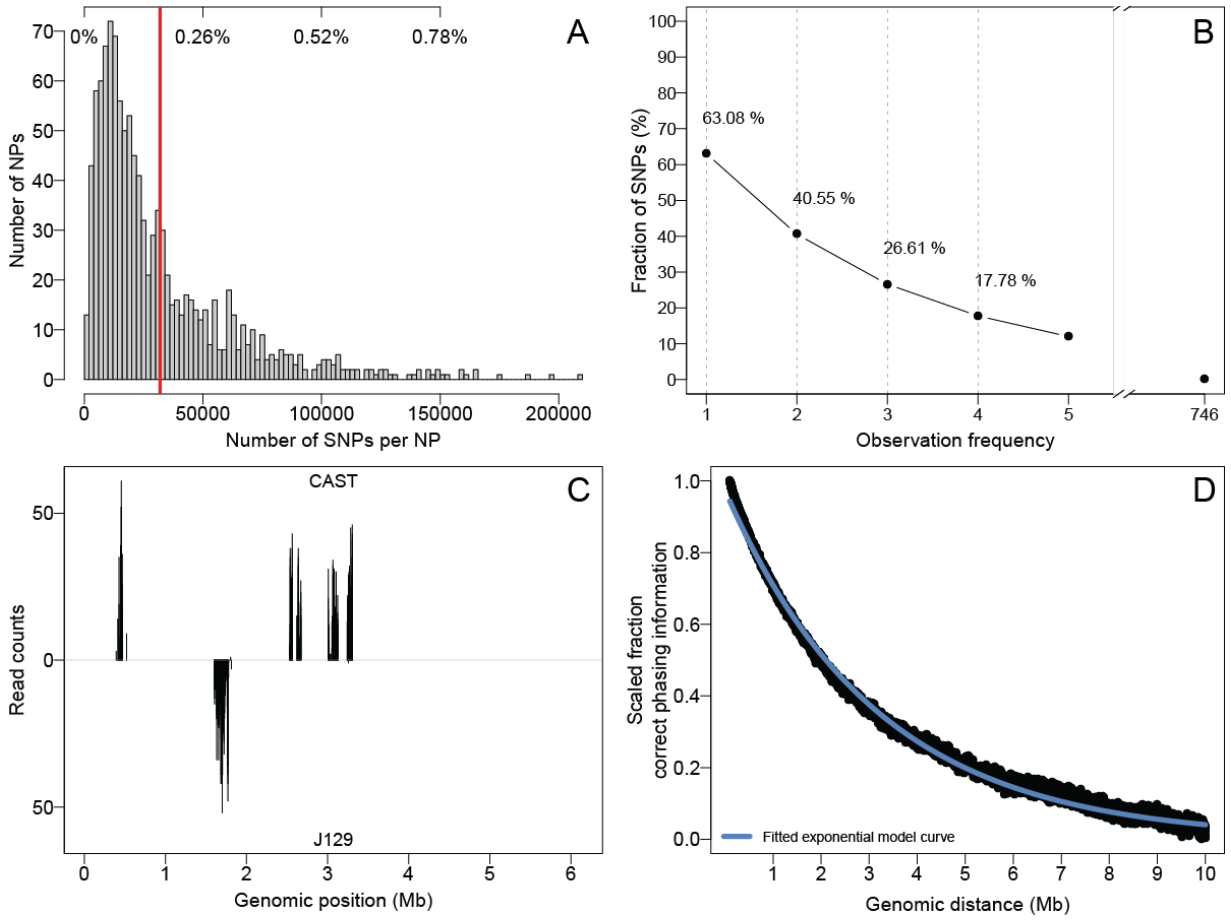

Figure 1: GAM captures local phasing information: A) Histogram of the number of observed SNPs per NP in the F123 dataset (fraction of all SNPs at top, mean = 0.166%, red line). **B)** Cumulative fraction of SNP observation frequencies. 63.08 % of SNPs are observed at least once, 40.55% of SNPs are observed at least twice across all NPs. **C)** Example of read counts supporting the CAST (upwards) and J129 (downwards) alleles in a single NP from chromosome 19. GAM captures small regions of the genome with local phasing information, where SNPs with small genomic distance co-observed in the same NP are most likely to originate from the same physical chromosome. **D)** The fraction of correct phasing information decreases with increasing linear distance of observed SNP pairs. The strongest decrease is apparent within 10 Mb genomic distance. The fit of the exponential model to the fraction of correct phasing information of SNP pairs with genomic distance between 100 kb and 10 Mb is shown in blue.

Haplotype reconstruction algorithms

Neighbour phasing:

Encouraged by the strong phasing signal of SNPs in immediate proximity to each other, we first considered a naive but fast phasing strategy as a reference that leverages the most reliable short-range haplotype information on neighbouring SNPs only (neighbour phasing). Let $N = 1123$ be the number of NPs and $K = 12,155,702$ be the number of neighbouring SNP pairs. Two possible relations of neighbouring SNP pairs are considered: a “flip” relation, where the alternative (alt) allele of one SNP is observed with the reference (ref) allele of the other SNP (alt-ref or ref-alt), and a “stay” relation, where both SNPs show either the reference or alternative allele (ref-ref or alt-alt). We build a $2 \times K$ incidence matrix, summing the observed number of stay and flip relations between all co-observed neighbouring SNPs across all NPs. A greedy

assignment of the most frequently observed SNP relations determines the inferred haplotype. Ties are broken by always assuming a "stay" relation.

Graph phasing:

We next extended the considered local proximity of SNPs from immediate neighbours to larger genomic windows using a graph-based approach (Figure 2). To improve efficiency each chromosome is segmented into windows of 20,000 SNPs with 10,000 SNPs overlap. Due to the high SNP density of the F123 genome, the average window of 20,000 observed SNPs spans 3,89 Mb in our dataset (range 110 kb - 13.38 Mb). For each window of $M = 20,000$ SNPs and $N = 1123$ NPs we create a $N \times M$ incidence matrix $I \in \{-1, 0, 1\}^{N \times M}$ which describes whether the reference ($I_{nm} = 1$), alternative ($I_{nm} = -1$), or no allele ($I_{nm} = 0$) of the SNP was observed. From I, the $M \times M$ adjacency matrix $A = I^t I$ is computed which contains the accumulated counts of the stay-flip relations summed over all NPs, such that positive values indicate more stay transitions ($I_{kl} > 0$: 'stay') and negative values indicate more flip transitions ($I_{kl} < 0$: 'flip'). An equal number of observed stays and flips leads to zero entries ($I_{kl} = 0$). To account for the exponential decay in phasing information with increasing genomic distance, the entries of A are optionally scaled by their predicted probability of coming from the same parental haplotype using our exponential model (proximity scaling, see above).

Non-zero entries in the adjacency matrix A induce an undirected weighted graph. The number of flip operations along a path between any two SNPs determines the haplotype assignment of their alternative alleles: if the number is even, both alleles reside on the same haplotype, if it is odd they reside on opposite haplotypes. Because multiple paths between SNPs can be conflicting in their haplotype assignment, we use the absolute value of the edge weights to compute a weighted maximum spanning tree (MST) using Kruskal's algorithm. This MST then defines a phasing graph representing the highest scoring haplotype structure per window.

The MSTs of overlapping windows are then joined into a chromosome-wide phasing graph G. A second iteration of the MST prunes the remaining low-scoring edges from G, creating the chromosome-wide phasing tree G'. Separate phasing blocks (separate potentially nested connected components of G') i and j, where the leftmost SNP of i precedes the leftmost SNP of j in genomic coordinates, are then merged such that the SNP with the leftmost position (in genomic coordinates) in phasing block j is connected via an assumed stay relation to the nearest preceding SNP in i.

To convert G' into a linear arrangement of alleles that defines a haplotype we choose the first SNP (leftmost SNP in genomic coordinates) as our anchor SNP and assign its alternative allele to haplotype A. We then count the number of flip operations on the path from the anchor SNP to any other SNP in G'. If this number is even, we assign the SNP to the same haplotype A as the anchor SNP, and otherwise to haplotype B.

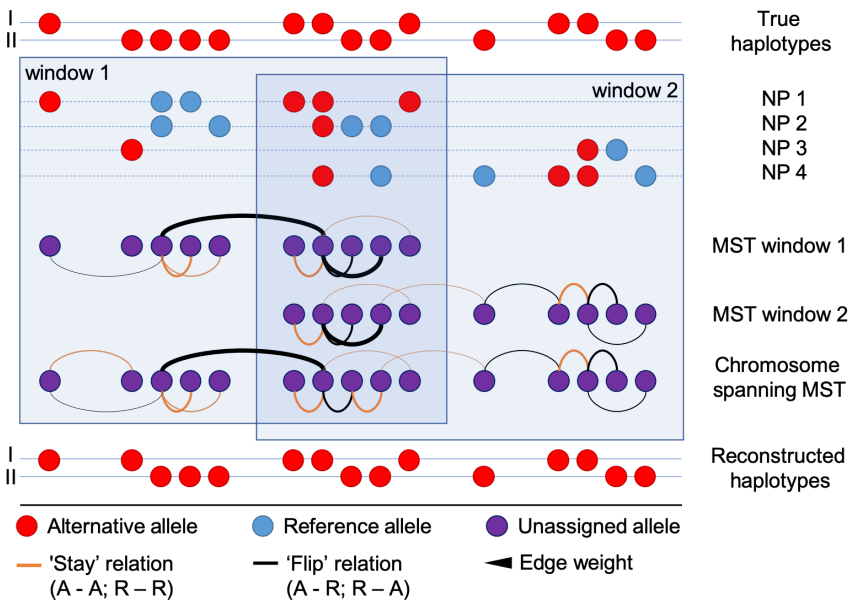

Figure 2: Schematic overview of the graph phasing algorithm. NPs (NP 1-4) are sparse local samples of the true haplotype structure (top). In overlapping windows, graphs of co-observed SNPs are built over all NPs. Edges are of either stay (orange) or flip (black) type and edge weights correspond to the co-observation frequency (line width) and are optionally proximity scaled. MSTs are calculated per window and combined to yield a chromosome-spanning MST. Unconnected blocks are concatenated by assumed stay relations. Finally, the chromosome-spanning MST is linearised to assign alternative alleles to the final reconstructed haplotypes.

Performance measures

We assess the performance of the haplotype reconstruction process relative to the 12,155,703 observed variants using two measures: (i) the correctness of the global haplotype structure, defined as a linear concatenation of alternative and reference alleles; (ii) the correctness of the local pairwise phase relationship between adjacent SNPs defined as the switch errors (Bansal 2019; Selvaraj et al. 2013; Geraci 2010). The global measure of allele assignments (i) is sensitive to individual switch errors, such that all alleles following a switch error will be considered wrongly assigned. The local measure (ii) circumvents this problem by focusing on the pairwise SNP relations only, but ignores overall haplotype correctness.

To assess reconstruction performance, we employ sensitivity, specificity and balanced accuracy (Brodersen et al. 2010), defining positive and negative classes as follows: for haplotype comparisons (i) reference alleles act as positive and alternative alleles as negative classes; for switch error comparisons (ii) stay transitions act as positive and flip transitions as negative classes. Balanced accuracy $BACC$ is a normalised variant of the correct classification rate adapted for unbalanced datasets:

$$BACC = \frac{1}{2}(Specificity + Sensitivity) = \frac{1}{2}\left(\frac{TN}{TN+TP} + \frac{TP}{TP+FN}\right)$$

We additionally consider the length of the largest phasing block, the total number of phasing blocks and the relative frequency of phased SNPs N_{phased} over all N variants ($f_{phased} = \frac{N_{phased}}{N}$) as measures of phasing completeness. To account for the different number of variants accessible to the neighbour and graph phasing approaches, we use balanced accuracy relative to the fraction of variants phased as a summary of local phasing performance.

GAMIBHEAR implementation

The presented haplotype reconstruction algorithms are implemented in the user-friendly R package GAMIBHEAR. It includes functions for parsing and cleaning of called variants from GAM experiments and different output functions next to the three phasing algorithms (neighbour phasing and graph phasing with or without proximity scaling). The user can visualize, process and compare intermediate results, restrict the analysis to target chromosomes or specific genomic regions, and apply custom filters such as individual quality cutoffs. The proximity-scaled graph phasing algorithm is time and memory efficiently implemented and parallelised to improve performance. GAMIBHEAR is open source and freely available under the GPL-2 license at https://bitbucket.org/schwarzlab/gamibhear.

Results

Accurate haplotype reconstruction from GAM nuclear profiles

We evaluated the performance of both neighbour and graph phasing algorithms with and without proximity scaling in relation to the number of NPs available, from 300 to all 1123 NPs, using the performance measures described above (Figure 3). We found that at least 300 samples were required to obtain reliable and robust phasing information.

Neighbour phasing performance

To assess the completeness of the predicted haplotype, we analyzed the number and size of phasing blocks in number of SNPs and genomic span. The neighbour phasing algorithm leads to multiple small unconnected phased blocks, on average 81,532 blocks per chromosome, with a median size of 5 SNPs (356 bp). The largest phasing block included 146 SNPs, the block with the largest genomic range spans 180,136 bp. Across all resulting blocks approximately 80.14 % (300 NPs) up to 83.39 % (1123 NPs) of the observed SNPs on all chromosomes were phased. However, the spanned genomic range in bp was significantly lower, with 27.59 % and 39.48 % of the genome phased in the 300 NP and 1123 NP datasets respectively.

Local performance using switch errors of all phasing blocks was high overall, showing 99.27 % sensitivity, 96.37 % specificity and a balanced accuracy of 97.82 % for the full dataset. We put the local balanced accuracy into perspective of the fraction of phased variants yielding a relative local balanced accuracy of 81.57 % (Figure 3A). Local performance was constantly high, independent from sample size. However, since the fraction of phased over observed SNPs increased slightly with increasing sample size, the relative local balanced accuracy increased as well. Due to the large number of unconnected phasing blocks, global performance evaluated through direct haplotype comparisons showed only 50.08 % sensitivity, 50.71 % specificity and a balanced accuracy of 50.39 % independent of sample size (Figure 3B). For an example of these locally accurately phased blocks reconstructed by the neighbour phasing algorithm see Figure 3C.

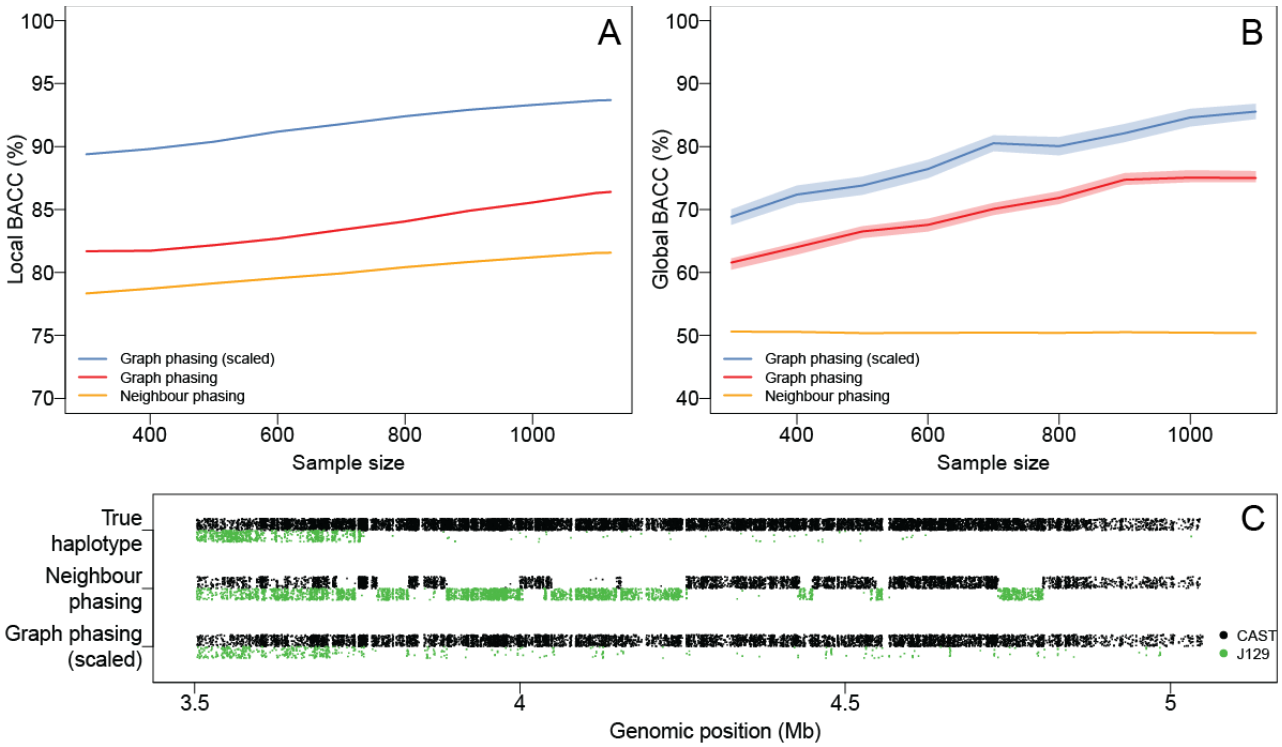

Figure 3: Performance results of reconstructed haplotypes using the neighbour-based algorithm (orange) and the basic (red) and proximity-scaled (blue) graph-based algorithms. The shaded areas indicate standard errors of the mean of 5 iterations of random sampling NPs from the dataset. **A)** Neighbour-phasing shows very high local balanced accuracy, however, scaled by ratio of phased variants it is outperformed by the graph-based algorithms. **B)** Due to the large number of unconnected phased blocks generated by the neighbour-phasing algorithm (see also C) its global balanced accuracy is only marginally better than random choice. Graph-based algorithms build one main chromosome spanning phased block which results in improved global balanced accuracy increasing with larger sample size. **C)** Example of SNP assignment on Mb 3.5-5 on chromosome 1; top lane depicts the true F123 haplotype of observed SNPs, alternative alleles on the CAST haplotype are colored in black, alternative alleles located on the J129 haplotype are shown in green. The middle track shows multiple phased blocks reconstructed by the neighbour phasing algorithm, which are locally highly accurate, but occasionally connected by incorrectly assumed phases between the blocks. The bottom track shows the haplotype reconstructed with the proximity-scaled graph phasing algorithm, showing improved global accuracy.

Graph phasing performance

Graph phasing created on average 100.58 (median 96) phasing blocks, the largest of which connects 99.91 % of observed variants and spans more than 98 % of the chromosome. The remaining blocks contain on average 2.7 SNPs (median 2 SNPs) and span on average 109,884 bp (median 31 bp). Notably, the size of the phasing blocks, the genomic span and the fraction of phased variants from observed was constant with increasing sample size beyond 300 NPs. Proximity scaling did not affect the completeness of the haplotype reconstruction.

As expected, the additional information leveraged by considering larger SNP windows compared to the neighbour phasing algorithm substantially improved global phasing accuracy to 71.43 % sensitivity, 73.26% specificity and a balanced accuracy of 72.35 % for the full dataset. Proximity scaling increased the global performance to 85.78 % sensitivity, 88.72 % specificity and a balanced accuracy of 87.25 % for the full dataset (Figure 3B). However, we observe most chromosomes to be phased with ~94 % global balanced accuracy, while a few suffer from single switch errors

flipping the predicted haplotype, reducing the respective global balanced accuracy to a minimum of 64%.
Local performance was high on the full dataset, yielding 88.60 % sensitivity, 84.38 % specificity and a balanced accuracy of 86.49 % without proximity scaling and 95.58 % sensitivity, 92 % specificity and a balanced accuracy of 93.79 % with proximity scaling of edge weights (Figure 3A). Both local and global phasing performance increased significantly with increasing sample size (Figure 3A-B). These results show that GAMIBHEAR provides reliable chromosome-wide haplotypes from Genome Architecture Mapping data. For an overview of the key performance metrics, see Table 1.

Table 1: Comparison of performance measures for neighbour-based phasing algorithm, basic and proximity-scaled graph phasing algorithm for the full dataset. Local balanced accuracy is additionally shown relative to the fraction of phased from observed SNPs.

	BACC local (%)	BACC global (%)	SNP phased (%)	BACC local (relative) (%)
Neighbour phasing	97.82	50.39	83.39	81.57
Graph phasing	86.49	72.35	99.95	86.45
Graph phasing (proximity-scaled)	93.79	87.25	99.95	93.74

Discussion

We presented the algorithm and software toolbox GAMIBHEAR for phasing of germline genetic variation from GAM nuclear slices. GAMIBHEAR is implemented as a user-friendly R package and is orthogonal to recently proposed computational approaches that exploit chromatin conformation data from Hi-C experiments for haplotype reconstruction (Selvaraj et al. 2013; Bansal 2019). In comparison to Selvaraj *et al.* (2013), the chromosome-spanning largest blocks resulting from GAMIBHEAR included >99.9% of variants compared to about ~95% of variants using HaploSeq. By combining a graph-based approach with a GAM-specific probabilistic model of chromosome capture we achieve high accuracy both in our global and local assessments of phasing performance. Within our probabilistic model we observed a stark dropoff in phasing information in more than 10 Mb distance from the source SNP. This drop off is likely due to the formation of highly interacting genomic regions and corresponding organisational chromatin structures such as self-interacting TADs (megabase scale) and higher order metaTADs which form depending on the transcriptional activity of the genomic region (Razin et al. 2016; Fraser et al. 2015; Ulianov et al. 2016).

Our proximity scaling model improves the haplotype reconstruction accuracy by not only assigning importance to variant relations based on the frequency of their observation, but also by taking genomic distances between variants into account. The MST through this proximity-scaled weighted graph reveals the most likely haplotype by discarding potential noise and assigning more importance to more likely co-observations of SNPs within neighbouring genomic regions. This approach runs the theoretical risk of breaking phasing blocks in situations where the only connecting variants were distant in genomic coordinates. In our dataset, no phasing blocks were broken due to proximity scaling of edge weights.

While GAMIBHEAR is ultimately intended to be used on human data, no GAM dataset of sufficient size is yet available on human samples. In the meantime, the F123 cell line is well-suited to accurately measure phasing performance due to its known haplotype structure before adapting the algorithm to the characteristics of human genomes. Traditionally, Hamming distance and Switch Error Rate have been widely used to measure phasing performance globally and locally (Bansal 2019; Selvaraj et al. 2013; Geraci 2010). However, in the F123 cell line 86.86 % of alternative alleles stem from the CAST parental genome. This imbalance in the distribution of alleles leads to inflated accuracy for these measures. Balanced accuracy (Brodersen et al. 2010) compensates for class imbalance by combining specificity and sensitivity, thus reflecting average correct classification rates for both classes. Completeness of the phasing, measured by the size (in number of covered SNPs) and the span (covered genomic range in bp) of the phasing blocks, is equally important. The "quality adjusted median size of the haploblocks" (QAN50) metric (Duitama et al. 2012) has been proposed in the past for this purpose. QAN50 does not accurately reflect phasing completeness of our graph phasing approach due to the strong skewness in the size distribution of its haploblocks. Thus we report the total number of phasing blocks and their size in number of covered SNPs and base pairs and choose relative balanced accuracy as a summary statistic.

Conclusion

Understanding the effect of genetic variation on chromatin conformation and gene regulation is a key question in genomics research. Large consortia, such as the 4D Nucleome project (Dekker et al. 2017), are now bundling resources to address this and allele-specific analyses of chromatin conformation and other sources of genomic variation are moving increasingly into the spotlight (Cavalli et al. 2019). The recently established GAM method (Beagrie et al. 2017) offers a unique opportunity towards high-resolution allele-specific analyses of chromatin contacts in humans, and GAMIBHEAR provides the necessary algorithmic advances towards generating highly accurate, chromosome-spanning haplotypes from GAM data on human samples in the future.

Author contributions

JM performed bioinformatic analysis on GAM data, developed and implemented the algorithms and R package. RK and GL produced the GAM data. AK generated the F123 reference genome. IIA and AK performed bioinformatic analysis and quality control of GAM data. RFS and AP designed and supervised the project.

Acknowledgements

The authors thank the Helmholtz Association (Germany) for support. AP acknowledges support from the National Institutes of Health Common Fund 4D Nucleome Program grant U54DK107977. IIA was supported by a Long-Term Fellowship from the Federation of European Biochemical Societies (FEBS). JM and RFS thank Birte Kehr and Sven Rahmann for valuable discussions.

References

Andrews, Simon. 2010. "FastQC - A Quality Control Tool for High Throughput Sequence Data." Babraham Bioinformatics. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bansal, Vikas. 2019. “Integrating Read-Based and Population-Based Phasing for Dense and Accurate Haplotyping of Individual Genomes.” *Bioinformatics* 35 (14): i242–48.

Bansal, Vikas, and Vineet Bafna. 2008. “HapCUT: An Efficient and Accurate Algorithm for the Haplotype Assembly Problem.” *Bioinformatics* 24 (16): i153–59.

Beagrie, Robert A., Antonio Scialdone, Markus Schueler, Dorothee C. A. Kraemer, Mita Chotalia, Sheila Q. Xie, Mariano Barbieri, et al. 2017. “Complex Multi-Enhancer Contacts Captured by Genome Architecture Mapping.” *Nature* 543 (7646): 519–24.

Brodersen, Kay Henning, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. “The Balanced Accuracy and Its Posterior Distribution.” *2010 20th International Conference on Pattern Recognition*. https://doi.org/10.1109/icpr.2010.764.

Browning, Sharon R., and Brian L. Browning. 2007. “Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering.” *The American Journal of Human Genetics*. https://doi.org/10.1086/521987.

Cavalli, Marco, Nicholas Baltzer, Husen M. Umer, Jan Grau, Ioana Lemnian, Gang Pan, Ola Wallerman, et al. 2019. “Allele Specific Chromatin Signals, 3D Interactions, and Motif Predictions for Immune and B Cell Related Diseases.” *Scientific Reports* 9 (1): 2695.

Chaisson, Mark J. P., Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, et al. 2019. “Multi-Platform Discovery of Haplotype-Resolved Structural Variation in Human Genomes.” *Nature Communications* 10 (1): 1784.

Dekker, Job, Andrew S. Belmont, Mitchell Guttman, Victor O. Leshyk, John T. Lis, Stavros Lomvardas, Leonid A. Mirny, et al. 2017. “The 4D Nucleome Project.” *Nature* 549 (7671): 219–26.

Duitama, Jorge, Gayle K. McEwen, Thomas Huebsch, Stefanie Palczewski, Sabrina Schulz, Kevin Verstrepen, Eun-Kyung Suk, and Margret R. Hoehe. 2012. “Fosmid-Based Whole Genome Haplotyping of a HapMap Trio Child: Evaluation of Single Individual Haplotyping Techniques.” *Nucleic Acids Research* 40 (5): 2041–53.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw354.

Fraser, James, Carmelo Ferrai, Andrea M. Chiariello, Markus Schueler, Tiago Rito, Giovanni Laudanno, Mariano Barbieri, et al. 2015. “Hierarchical Folding and Reorganization of Chromosomes Are Linked to Transcriptional Changes in Cellular Differentiation.” *Molecular Systems Biology* 11 (12): 852.

Geraci, Filippo. 2010. “A Comparison of Several Algorithms for the Single Individual SNP Haplotyping Reconstruction Problem.” *Bioinformatics* 26 (18): 2217–25.

Goios, Ana, Luísa Pereira, Molly Bogue, Vincent Macaulay, and António Amorim. 2007. “mtDNA Phylogeny and Evolution of Laboratory Mouse Strains.” *Genome Research* 17 (3): 293–98.

Gribnau, Joost, Konrad Hochedlinger, Ken Hata, En Li, and Rudolf Jaenisch. 2003. “Asynchronous Replication Timing of Imprinted Loci Is Independent of DNA Methylation, but Consistent with Differential Subnuclear Localization.” *Genes & Development* 17 (6): 759–73.

Jamal-Hanjani, M., and G. A. Wilson. 2017. “Tracking the Evolution of Non–small-Cell Lung Cancer.” *England Journal of ….* https://www.nejm.org/doi/full/10.1056/NEJMoa1616288.

Jun, Goo, Mary Kate Wing, Gonçalo R. Abecasis, and Hyun Min Kang. 2015. “An Efficient and Scalable Analysis Framework for Variant Extraction and Refinement from Population-Scale DNA Sequence Data.” *Genome Research* 25 (6): 918–25.

Kempfer, Rieke, and Ana Pombo. 2019. “Methods for Mapping 3D Chromosome Architecture.” *Nature Reviews. Genetics*, December. https://doi.org/10.1038/s41576-019-0195-2.

Khalil, A., J. L. Grant, L. B. Caddle, E. Atzema, K. D. Mills, and A. Arneodo. 2007. “Chromosome Territories Have a Highly Nonspherical Morphology and Nonrandom Positioning.” *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 15 (7): 899–916.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding

Principles of the Human Genome." *Science* 326 (5950): 289–93.
Li, Heng. 2011. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data." *Bioinformatics* 27 (21): 2987–93.
Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics* 25 (14): 1754–60.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp352.
Loh, Po-Ru, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, et al. 2016. "Reference-Based Phasing Using the Haplotype Reference Consortium Panel." *Nature Genetics* 48 (11): 1443–48.
Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.journal* 17 (1): 10–12.
Meaburn, Karen J., and Tom Misteli. 2007. "Cell Biology: Chromosome Territories." *Nature* 445 (7126): 379–781.
O'Sullivan, Justin M., Michael D. Hendy, Tatyana Pichugina, Graeme C. Wake, and Jörg Langowski. 2013. "The Statistical-Mechanics of Chromosome Conformation Capture." *Nucleus* 4 (5): 390–98.
Patterson, Murray, Tobias Marschall, Nadia Pisanti, Leo van Iersel, Leen Stougie, Gunnar W. Klau, and Alexander Schönhuth. 2015. "WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 22 (6): 498–509.
PCAWG Transcriptome Core Group, Claudia Calabrese, Natalie R. Davidson, Deniz Demircioğl, Nuno A. Fonseca, Yao He, André Kahles, et al. 2020. "Genomic Basis for RNA Alterations in Cancer." *Nature*, 1–8.
Razin, Sergey V., Alexey A. Gavrilov, Yegor S. Vassetzky, and Sergey V. Ulianov. 2016. "Topologically-Associating Domains: Gene Warehouses Adapted to Serve Transcriptional Regulation." *Transcription* 7 (3): 84–90.
Selvaraj, Siddarth, Jesse R Dixon, Vikas Bansal, and Bing Ren. 2013. "Whole-Genome Haplotype Reconstruction Using Proximity-Ligation and Shotgun Sequencing." *Nature Biotechnology* 31 (12): 1111–18.
Simpson, E. M., C. C. Linder, E. E. Sargent, M. T. Davisson, L. E. Mobraaten, and J. J. Sharp. 1997. "Genetic Variation among 129 Substrains and Its Importance for Targeted Mutagenesis in Mice." *Nature Genetics* 16 (1): 19–27.
Tewhey, Ryan, Vikas Bansal, Ali Torkamani, Eric J. Topol, and Nicholas J. Schork. 2011. "The Importance of Phase Information for Human Genomics." *Nature Reviews. Genetics* 12 (3): 215–23.
Ulianov, Sergey V., Ekaterina E. Khrameeva, Alexey A. Gavrilov, Ilya M. Flyamer, Pavel Kos, Elena A. Mikhaleva, Aleksey A. Penin, et al. 2016. "Active Chromatin and Transcription Play a Key Role in Chromosome Partitioning into Topologically Associating Domains." *Genome Research* 26 (1): 70–84.