bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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], Birte Kehr[3,4], Ana Pombo[1,2], Sven Rahmann[5], Roland F Schwarz[1,#]

[1] Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Hannoversche Str. 28, 10115 Berlin, Germany
[2] Department of Biology, Humboldt University of Berlin, Unter den Linden 6, 10099 Berlin, Germany
[3] Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany
[4] Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Charitéplatz 1, 10117 Berlin, Germany
[5] Genome Informatics, Institute of Human Genetics, University Hospital Essen, University of Duisburg-Essen, 45122 Essen, Germany

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

Abstract

Motivation

Genome Architecture Mapping (GAM) was recently introduced as a digestion- and ligation-free method to detect chromatin conformation. Orthogonal to existing approaches based on chromatin conformation capture (3C), GAM's ability to capture both inter- and intra-chromosomal contacts from low amounts of input data makes it particularly well suited for allele-specific analyses in a clinical setting. Allele-specific analyses are powerful tools to investigate the effects of genetic variants on many cellular phenotypes including chromatin conformation, but require the haplotypes of the individuals under study to be known a-priori. So far however, no algorithm exists for haplotype reconstruction and phasing of genetic variants from GAM data, hindering the allele-specific analysis of chromatin contact points in non-model organisms or individuals with unknown haplotypes.

Results

We present GAMIBHEAR, a tool for accurate haplotype reconstruction from GAM data. GAMIBHEAR aggregates allelic co-observation frequencies from GAM data 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 GAMIBHEAR to infer accurate genome-wide haplotypes from GAM data.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 Introduction

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 (NuPs) (Beagrie et al. 2017). NuPs are generated through cryosectioning of cellular nuclei followed by next-generation sequencing. Chromatin contacts between DNA loci can be inferred by analysing the frequency at which the loci are captured in the same NuP. In contrast to ligation-based chromatin conformation capture (3C) type approaches, such as Hi-C (Lieberman-Aiden et al. 2009), GAM is able to resolve complex contacts with three or more loci with high resolution, does not suffer from non-uniformity biases (Chandradoss et al. 2020) and only requires several hundreds of cells to obtain high-resolution contact maps (Kempfer and Pombo 2019; Beagrie et al. 2020; Fiorillo et al. 2020). This makes GAM particularly useful for the study of chromatin contacts in rare biological materials, such as human biopsies.

Recently, there has been increasing interest in the allele-specific analysis of chromatin contacts, for which haplotyping, i.e. phasing of single nucleotide variants (SNVs) is key (Chen et al. 2017; Rivera-Mulia et al. 2018; Cavalli et al. 2019; Zahn 2020). Traditionally, haplotypes are inferred through read-based phasing methods such as HapCut and WhatsHap (Bansal and Bafna 2008; Patterson et al. 2015; Edge, Bafna, and Bansal 2017) or statistically using population-level or reference-phasing approaches such as SHAPEIT and BEAGLE (Loh et al. 2016; Browning and Browning 2007). Read-based haplotype phasing can be formalised in a number of ways. Variants of the Minimum Error Correction (MEC) problem have frequently been used in the presence of different error distributions and insert lengths (Bansal and Bafna 2008). MEC views the given data (a fragments by SNV sites matrix of observed allele states) as potentially erroneous and asks for the least invasive way to correct the observations in order to enable conflict-free phasing. The MEC problem has been demonstrated to be computationally hard under a variety of conditions (Bafna et al. 2005).

Initial efficient MEC heuristics for short-read sequencing data such as HapCut, which converts MEC to a minimum cut problem, only allowed for single base pair errors (Bansal and Bafna 2008). Motivated by observations that homologous chromosomes tend to occupy distant chromosome territories (Meaburn and Misteli 2007), Selvaraj et al. (2013) proposed HaploSeq to leverage Hi-C data with an extension of the HapCut algorithm to accommodate Hi-C specific h-trans errors. H-trans errors are haplotype switch errors that occur when a piece of DNA interacts with a DNA fragment from the homologous chromosome rather than the same chromosome. HapCut2, which was recently released, includes population-based statistical phasing (Bansal 2019) and implements a variety of different error models to accommodate different sequencing technologies (Edge, Bafna, and Bansal 2017).

Another approach that yields an NP-hard minimum cut problem seeks to partition the observed fragments into two classes corresponding to the two haplotypes, again by minimising a measure of inconsistency (Duitama et al. 2010). Other methods first use an aggregation step to collect co-occurrence frequency evidence of SNVs at different positions and then seek to reconstruct one of the two haplotypes by partitioning the SNV sites into two classes; this can be written as a well-known problem from physics: finding a ground state of a spin glass system, which is also a minimum cut problem (Tourdot and Zhang 2019).

We here ask the question to what degree GAM data can be used effectively for haplotyping. The coverage and error distributions of the GAM cryosectioning process are sufficiently different from Hi-C based approaches that existing MEC solvers are not directly applicable. Hi-C data yields ligated

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reads of genomic loci which can be very distant in linear genomic space but typically from the same chromosomal haplotype. In contrast, most GAM NuPs yield individual short reads of both haplotypes and only maintain haplotype fidelity locally. Thus, in contrast to Hi-C, where h-trans errors remain rare, GAM NuPs frequently switch haplotypes. In addition, SNV coverage in GAM data varies greatly and non-uniformly, which interferes with MEC solvers that require the maximum coverage per SNV to be low (Patterson et al. 2015).

To explore the potential of GAM for haplotyping, we present GAMIBHEAR (GAM-Incidence Based Haplotype Estimation And Reconstruction), a novel computational tool for whole-genome phasing of genetic variants from GAM NuPs. Similar to previous haplotyping approaches for other types of data, we use an aggregation step and formulate the problem on co-occurrence evidence derived from the raw GAM data (see Sec. 2.1). GAMIBHEAR then employs a graph representation of the co-occurence of SNV alleles in NuPs 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. The formulation is similar to the above mentioned physics problem of finding a ground state of a spin glass system (Tourdot and Zhang 2019). We assess the performance of GAMIBHEAR on the hybrid mouse embryonic stem cell line F123 with known haplotype structure. Despite the sparsity of GAM data, GAMIBHEAR allows for accurate long-distance haplotype reconstruction. GAMIBHEAR is available as an efficient R package with parallel implementations of the most compute-intensive tasks and is available at https://bitbucket.org/schwarzlab/gamibhear.

2 Methods

2.1 Definitions, problem statement and objective

Our goal is to reconstruct haplotypes from GAM data. A sequenced GAM dataset consists of reads from many nuclear profiles (NuPs). Each NuP 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 but not limited to loci proximal in linear distance. Thus, reads from single NuPs cover a small proportion of the whole genome with consecutive stretches of genomic DNA that reflect chromatin looping in and out of a thin nuclear slice (illustrated in Fig. 2B). Our underlying assumption for haplotype reconstruction is that alleles of any two heterozygous SNVs captured in a nuclear slice are likely to originate from the same parental copy, and that this likelihood decreases with increasing genomic distance of the co-observed alleles.

We assume that the set of heterozygous SNVs is given and that the SNV alleles have been determined per NuP. The input data to the GAM haplotype reconstruction problem can thus be described as follows: Let N be the number of NuPs and M be the number of heterozygous SNVs in the genomic region of interest (e.g., a chromosome or chromosome arm; sites with homozygous SNVs are ignored). Then the problem input is a ternary $N \times M$ matrix D with $D_{ij} = 1$ if the reference allele is observed in NuP i at SNV site j, $D_{ij} = -1$if the alternative allele is observed, and $D_{ij} = 0$if there is no unique observation (e.g. due to lack of coverage or if both alleles are observed in the same NuP).

The goal is to reconstruct the two haplotypes (allele states on the same parental copy). Formally, a haplotype is a vector $h \in \{-1,1\}^M$with $h_j = 1$ if the reference allele is found at site j and $h_j = -1$for the alternative allele. One of the two haplotypes h determines the other one as $-h$.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The GAM input data in principle contains the information to infer$h$. Consider the relation between SNV sites j and k in NuP i. The two sites can be in a "flip" relation, where the alternative (alt) allele (-1) of one site is observed with the reference allele (+1) of the other site (product$D_{ij} \cdot D_{ik} = -1$), and a "stay" relation, where both SNVs show either the reference or alternative allele (product $D_{ij} \cdot D_{ik} = 1$).
We thus compute the $M \times M$ *evidence matrix* $A := D^T D$, which contains the accumulated counts of the stay-flip relations summed over all NuPs, i.e. $A_{jk} = \sum_{i=1}^{N} D_{ij} \cdot D_{ik}$, such that positive values indicate more stay observations ($A_{jk} > 0$: 'stay' between sites j and k; $j, k = 1, \dots, M$) and negative values indicate more flip observations ($A_{jk} < 0$: 'flip' between sites j and k). An equal number of observed stays and flips leads to zero entries ($A_{jk} = 0$). In principle, we can additionally introduce NuP-specific reliability weights $\lambda = (\lambda_i) > 0$ with mean 1, and more generally define$A := D^T \Lambda D$, where Λ is the $N \times N$ diagonal matrix containing the λ_i.

The goal of the haplotype reconstruction algorithms we develop here is to solve h using the information contained in A: If $A_{jk} > 0$, then we should have $h_j = h_k$, and if $A_{jk} < 0$, then $h_j = -h_k$. However, the information in A may be conflicting when considering transitivity: Consider three sites j, k, l with $A_{jk} > 0, A_{kl} > 0, A_{jl} < 0$. Thus, decisions need to be made on how to resolve conflicting information in the evidence matrix A.

One possible formulation of the problem is as follows: Given the $M \times M$ matrix A, we seek $h \in \{-1,1\}^M$ to

$$\text{maximise } F(h) := \sum_{j<k} h_j A_{jk} h_k = h^T A h.$$

This formulation encourages h_j and h_k to take the same sign if $A_{jk} > 0$ and different signs if $A_{jk} < 0$. This maximization problem is equivalent to finding an exact ground state for a spin glass in physics and is known to be NP-hard in general and can be cast as a minimum cut problem on a graph induced by A (Tourdot and Zhang 2019). Here we propose heuristic algorithms that make use of known properties of the evidence matrix A (potentially proximity-scaled; see below) and evaluate them against a dataset with a known correct solution.

Before we state two such algorithms, let us first relax our notion of what we accept as a solution. Above, we defined a (fully resolved) haplotype as a vector $h \in \{-1,1\}^M$with $h_j = 1$ if the reference allele is found at site j and $h_j = -1$for the alternative allele. However, the available data may not be sufficient to fully resolve the haplotype. Instead of guessing, we allow partial solutions ("blocks") as follows. Let $\mathcal{J} := (J_1, J_2, \dots, J_K)$ be a partition (disjoint union) of $\{1, \dots, M\}$into K blocks. Then a solution of the GAM haplotype reconstruction problem for input matrix D with partition $\mathcal{J}$ is a collection of K binary vectors $h^1 \in \{-1,1\}^{J_1}, \dots, h^K \in \{-1,1\}^{J_K}$. Each of the K blocks is solved independently, and no statement is made about the connection between these blocks. The blocks are often intervals, but may be arbitrary subsets of all sites, especially for GAM data. Obviously, solutions with fewer independent blocks are more desirable.

2.2 Haplotype reconstruction algorithms

2.2.1 Neighbour phasing

We first consider a baseline phasing strategy that leverages the most reliable short-range haplotype information on neighbouring SNVs only ("neighbour phasing"). In the above notation, we only consider the first off-diagonal of A, i.e., $A_{j,(j+1)}$for $j = 1, \dots, M$. Essentially, this resolves possible conflicting information by ignoring a large fraction of the available data, and only considering a

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

single path between any two sites $j \leq k$: $j \rightarrow j+1 \rightarrow \cdots \rightarrow k$. The reconstructed haplotype starts (arbitrarily) with the reference allele, thus $h_1 = 1$. Once h_j is determined, we set $h_{j+1} := h_j \cdot \text{sign}(A_{j,(,j+1)})$, i.e. we stay or flip according to the sign of $A_{j,(j+1)}$. In case of a tie or when SNV *j* and *j+1* are never co-observed in the same NuP ($A_{j,(j+1)} = 0$), we start a new independent block where $h_{j+1} = 1$. Solutions produced by neighbour phasing consist of blocks that are intervals. The resolved blocks can be expected to be correct with high probability, but also short, and therefore of limited use.

2.2.2 Graph phasing with optional proximity scaling

We extend the considered local proximity of SNVs from immediate neighbours to larger genomic windows using a graph-based approach (Figure 1). To improve efficiency, each chromosome is segmented into windows of a fixed number L of SNV sites with half a window size overlap. To process a window, we restrict the $N \times M$ input matrix $D = (D_{ij})$ to the window's sites and only consider the reduced $N \times L$ matrix D and the derived $L \times L$ evidence matrix $A = (A_{jk})$.

As we assume that the reliability of phasing information within a NuP decreases with genomic distance, we include an option to scale the information in A element-wise by a weight matrix $W = (W_{jk})$, where W_{jk} depends on the genomic distance d_{jk} between sites j and k. We use a simple exponential decay model, where $W_{jk} = C \cdot exp\ (-\lambda\ d_{jk})$for d_{jk} in a certain range $[D_{min}, D_{max}]$, and $W_{jk} = 1$ for $d_{jk} < D_{min}$ and $W_{jk} = 0$for $d_{jk} > D_{max}$. The choice of appropriate parameters $C > 0, \lambda > 0$ and $0 \leq D_{min} < D_{max}$ is discussed below. In the following, A represents the *proximity-scaled evidence matrix* ($A_{jk} \leftarrow W_{jk} \cdot A_{jk}$).

At this point, there are four potential reasons for $A_{jk} = 0$: First, sites j and k may never co-occur in any NuP. Second, they may never be considered in the same window of L sites. Third, their genomic distance may be larger than D_{max}. Fourth, an equal number of observations of stay and flip relations may be encountered between sites $j$and k.

The non-zero entries in A induce an undirected weighted graph. Its L vertices are the sites of the current window. An edge between sites j and k exists with weight A_{jk} if $A_{jk} \neq 0$. Two sites in the same connected component of this graph are typically connected by many paths. Consider a single arbitrary path between sites j and k. The number of negative-weighted edges along the path determines the haplotype assignment: if the number is even, then $h_k = h_j$; if it is odd, then$h_k = -h_j$. Different paths between the two sites can be conflicting in their haplotype assignment. However, if the graph is reduced to a tree (or forest in case of more than one connected component), there is a unique path between each pair of sites (in the same connected component). Because the absolute values $|A_{jk}|$ indicate strength of direct evidence for the flip or stay operation between sites $j$and k, we compute a *maximum spanning tree* (MaxST) of each connected component based on absolute edge weights $|A_{jk}|$. This is done by Kruskal's algorithm, which is typically used to compute a minimum spanning tree in $O(m\ log\ n)$ time for a graph with m edges and n vertices, but can also be used to find a MaxST by negating the weights. Recall that the problem is solved on (potentially dense graphs of) windows, so the required running time is $O(L^2\ log\ L)$for each window. The MaxST approach has the property that the resulting path between any two sites j and k maximises the minimum weight of the path's edges among all possible paths between j and k (Hu 1961), so we construct the graph by maximising the weakest evidence link between each pair of sites of the window, which appears to be a reasonable heuristic for the given problem. The computed MaxST then determines the haplotypes (or set of haplotype blocks in case of a forest of MaxSTs) for the current window.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To infer haplotypes across the whole chromosome, the MaxSTs of overlapping windows must then be joined into a chromosome-wide tree. For this, we join the (overlapping) MaxSTs of all windows into a new graph consisting of all M SNV sites as nodes and the union of edges of all MaxSTs. Because each node is in at most two MaxSTs, the number of edges in the union is bounded by $2(M-1)$. For this sparse graph, we again determine a MaxST (if necessary, on each connected component separately) in $O(M\ log\ M)$ time to obtain a unique path between any two connected sites.

For the output, each connected component defines an independent block. We arbitrarily set the haplotype state of leftmost SNV site h_1 (with smallest genomic coordinate) in each block to $h_1 := 1$ (alternative allele), and compute the other states h_j according to the number of negative-weighted edges on the unique MaxST path between the first site and j.

Including phasing information from non-adjacent SNV pairs will improve completeness and yield larger, potentially chromosome-spanning haplotype blocks. In the reconstructed haplotypes of the graph phasing approach, blocks can be nested. The inclusion of phasing information from more distant SNV pairs might compromise the overall accuracy of the results, however the proximity scaling is expected to keep the introduction of misleading information to a minimum.

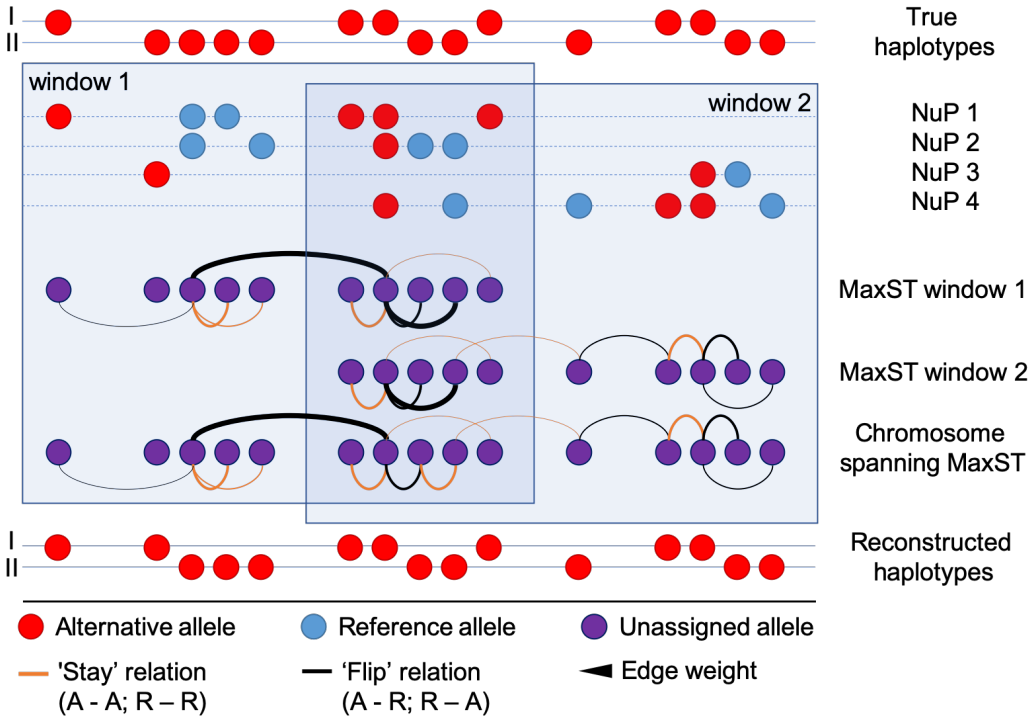

Figure 1: Schematic overview of the graph phasing algorithm. The location of alternative alleles of heterozygous SNVs on the two parental chromosomes describes the true haplotypes (top). NuPs 1-4 are sparse local samples of the true haplotype structure. In overlapping windows, graphs of co-observed SNVs are built over all NuPs. 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. MaxSTs are calculated per window and combined to yield a chromosome-spanning MaxST. Finally, the chromosome-spanning MaxST is used to assign alternative alleles to the final reconstructed haplotypes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.3 Performance measures

The overall quality of reconstructed haplotypes depends on both the completeness of the reconstructed haplotype blocks as well as the phasing accuracy of the SNVs contained. In addition to the total proportion of heterozygous SNVs that have been phased, metrics of completeness and contiguity assess the genomic distances spanned by the reconstructed haplotype blocks. Metrics proposed for this purpose include the N50 (Lander et al. 2001) , S50 (Lo et al. 2011) and AN50 (Lo et al. 2011) metrics. Briefly, in N50, the phased haplotype blocks are sorted by decreasing span (in base pairs; bp), and the span of the block at which 50% of variants are phased is determined. Analogously, the S50 metric uses the number of SNVs phased within the blocks instead of the genomic span in bp to determine the 50% phased threshold, thus accounting for species-specific SNV densities in genomes. To enable comparisons with previous investigations into the F123 cell line (Selvaraj et al. 2013) we report N50 as percent of the phasable genome (range between leftmost SNV and rightmost SNV per chromosome) and S50 as percent of phasable variants (number of input variants). We additionally report the adjusted N50 (AN50), which corrects the N50 measure for cases where smaller isolated haplotype blocks are contained within blocks spanning them, a typical scenario in graph-based phasing algorithms. Thus the genomic span is adjusted by the fraction of SNV phased in the range of the respective block.

To assess the accuracy of the reconstructed haplotypes we compare GAMIBHEAR estimates with the haplotypes of the F123 mouse embryonic stem cell (mESC) line obtained from whole-genome sequencing of the parental mouse strains (see Supplementary Note S1 'Benchmark genome (F123)'). Two measures are considered: First, the Switch Error Rate (SER), defined as the proportion of adjacent variant pairs that were phased incorrectly out of all phased variant pairs. The SER metric can be adjusted for highly fragmented results by introducing a penalty of 0.5 switch errors per unphased transition of neighbouring SNVs. The second measure is the global haplotype agreement calculated by direct comparison of the reconstructed and true haplotypes (i.e. alt-ref configurations) within haplotype blocks. SER is a more lenient metric compared to global haplotype accuracy, as a single switch error in the middle of a haplotype block will lead to half the haplotype block being assigned to the opposite haplotype.

To evaluate how the quality of the reconstructed haplotypes depends on the number of available NuPs, we reconstructed haplotypes using different sample sizes. In 10 iterations each, increasing numbers of NuPs were randomly sampled from the full dataset and haplotypes were reconstructed from the subsampled datasets.

All metrics are calculated per chromosome and the mean value and standard deviation over all chromosomes are reported.

2.4 GAMIBHEAR implementation

The presented haplotype reconstruction algorithms are implemented in the R package GAMIBHEAR. The package includes functions for parsing and cleaning of called variants from GAM experiments and different output functions in addition to the two phasing algorithms (neighbour phasing and graph phasing with optional proximity scaling). The user can visualise, process and compare intermediate results, restrict the analysis to target chromosomes or specific genomic regions, and apply custom filters such as individual quality cut-offs. The basic and 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3 Results

3.1 Dataset Statistics

Benchmark genome (F123). The F123 mouse embryonic stem cell line was derived from a hybrid F1 mouse resulting from the cross of the two inbred, homozygous mouse strains CAST (*Mus musculus castaneus*) and J129 (*Mus musculus domesticus J129*). The parental mouse strains are both fully sequenced, their exclusively homozygous genomic variants with respect to the reference mouse genome mm10, which was derived from the mouse strain C57BL/6, are known. The F1 generation resulting from the cross of CAST and J129 is thus heterozygous at al loci for which their parents have different alleles. Their haplotypes are known, making them an ideal model for benchmarking phasing algorithms. 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 C57BL/6, 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 SNV positions and 1,119,044 SNVs located in genomic regions of low mappability, the F123 reference set contains 18,150,228 variants in total, all of which are heterozygous due to inbreeding of the parental strains. This yields an average SNV density of 1 SNV per 132bp, with a median genomic distance of 56 bp.

Nuclear profiles. We obtained 1261 GAM NuPs of the F123 mESC cell line (4D Nucleome Consortium data portal accession number 4DNBSTO156AZ), out of which 1123 passed quality screening (see Supplementary Note S2).
We extracted on average 305,377 reads from each NuP, covering 0.171% (± 0.167) of the 18,150,228 heterozygous SNVs per nuclear slice (Figure 2A); exemplary data of genomic regions captured in a single NuP is shown in Figure 2B. Out of all F123 SNVs, 11,741,055 (64.69%) were observed at least once across all 1123 NuPs and 7,605,321 SNVs (41.9%) were observed at least twice (Figure 2C). Due to this sparsity and the fact that homologous chromosome pairs occupy distinct chromosomal territories (Khalil et al. 2007), 96.54% of SNV observations showed counts from only one parental allele within one sample. Thus, we removed observed variants with read counts from both parental alleles without substantial loss of information. Since the slicing of nuclei in the GAM experiments is a random process, a balanced observation ratio of alternative and reference alleles of heterozygous SNVs is expected. To minimise possibly confounding observations, statistically mono-allelic observations of SNVs were excluded from the dataset (binomial test against 0.5, p-values corrected for multiple testing).

3.2 Exponential proximity scaling

Our method includes the option of exponentially downweighting evidence information A_{jk} with increasing genomic distance (see Section 2.2.2). This raises the question whether indeed alleles of any two SNVs captured in a nuclear slice are more likely to originate from the same parental copy and whether this probability decreases with increasing genomic distance of the co-observed alleles. With the benchmark dataset with known haplotypes available, we were able to examine the empirical probability p of two alleles coming from the same haplotype based on their genomic distance d and fit an exponential function $p = C \cdot e^{-\lambda \cdot (d - D_{min})}$ using non-linear least squares. For this model we only considered pairs of sites within the interval $[D_{min}, D_{max}]$= [1 bp, 10 Mbp], where the decay in phasing information is most pronounced (Figure 2D). For pairs outside that distance range, which can be individually assigned by the user, probabilities 1 and 0 were assumed, respectively. Parameter $C = 1$ then describes the co-observation probability at a genomic distance

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of 1 bp with an exponential decay parameter of $\lambda = 3.173 \cdot 10^{-7}$. The simple exponential dependency well describes the empirical distribution (Figure 2D) and thus appears to be a good model for the reliability of the raw evidence as a function of genomic distance. In the following, we evaluate our graph phasing approach with and without proximity scaling.

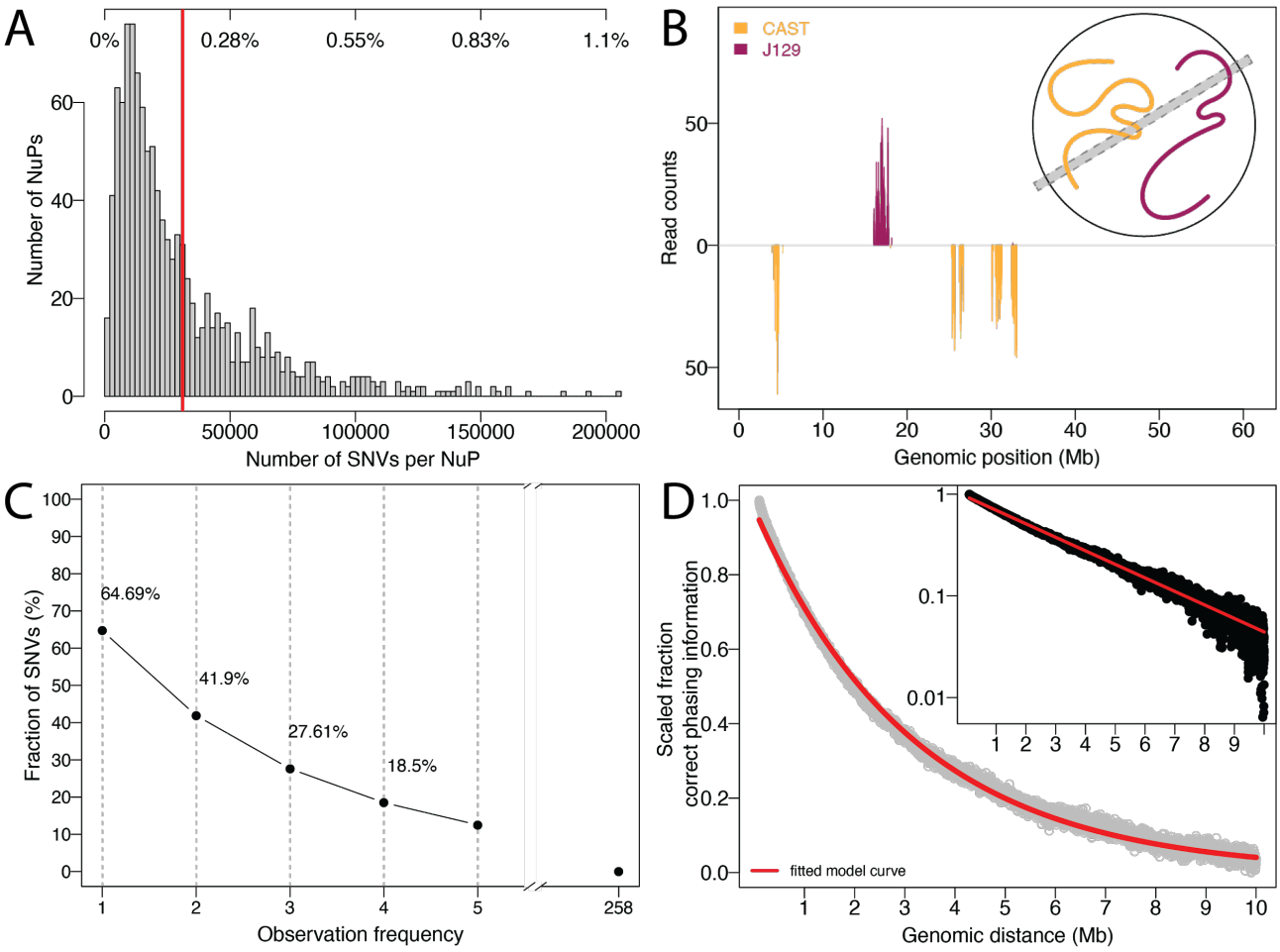

Figure 2: GAM captures local phasing information: A) Histogram of the number of observed SNVs per NuP in the F123 dataset (fraction of all SNVs at top, mean = 0.171%, red line). **B)** Example of read counts supporting the CAST (orange, downwards) and J129 (red, upwards) alleles in a single NuP on chromosome 19, visualising the sparsity of GAM data. Inset schematises physical capturing of respective genomic regions in a slice (grey area) by cryosectioning in a GAM experiment. **C)** Cumulative fraction of SNV observation frequencies. 64.69% of SNVs are observed at least once, 41.9% of SNVs are observed at least twice across all NuPs. **D)** The fraction of correct phasing information decreases exponentially with increasing genomic distance of observed SNV pairs. The fit of the exponential curve to the fraction of correct phasing information of SNV pairs with genomic distance between 1 bp and 10 Mbp is shown in red. The inset shows the decrease of correct phasing information on a logarithmic scale.

3.3 Performance of GAMIBHEAR

3.3.1 High quality haplotype reconstruction from GAM samples

We evaluated the quality of the haplotypes reconstructed with GAMIBHEAR in terms of completeness and accuracy using the true haplotypes of the F123 cell line, which were generated

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from the known genotypes of the parental strains (see Supplementary Note S1). A summary of the following scores can be found in Table 1.

Neighbour phasing performance. The neighbour phasing algorithm was built to exploit the most reliable short-range haplotype information of neighbouring co-observed SNVs, at the expense of completeness. The low switch error rate of 0.76% (±0.13%) (Figure 3A) of the haplotypes reconstructed with the strict neighbour phasing algorithm demonstrates strong local phasing information in GAM NuPs. However, adjusting the SER by penalising each unphased transition between adjacent SNVs by 0.5 switch errors (Figure 3A) results in a mean adjusted SER of 6.61% (±0.18%). Due to the sparsity of the GAM data (Figure 2B), neighbour phasing completeness is low (Figure 3C) and neighbour phasing cannot generate chromosome-spanning haplotypes. Although 95.94% (±0.25%) of input SNVs were phased into haplotypes blocks of size 2 or larger, due to the large number of gaps between blocks only 83.02% (±0.58%) of possible transitions between neighbouring SNV pairs could be phased (Figure 3D). Half of the SNVs were phased in blocks connecting less than 11 SNVs (± 1) (S50) and spanning less than 742 bp (±41bp) (N50, Figure 3C). Since this algorithm does not allow for blocks spanning unphased variants or nested blocks, N50 does not need to be adjusted. As each small block can be placed arbitrarily on either of the two haplotypes, a global measure of haplotype accuracy is uninformative and we thus omitted neighbour phasing from the global accuracy measure (Figure 3B).

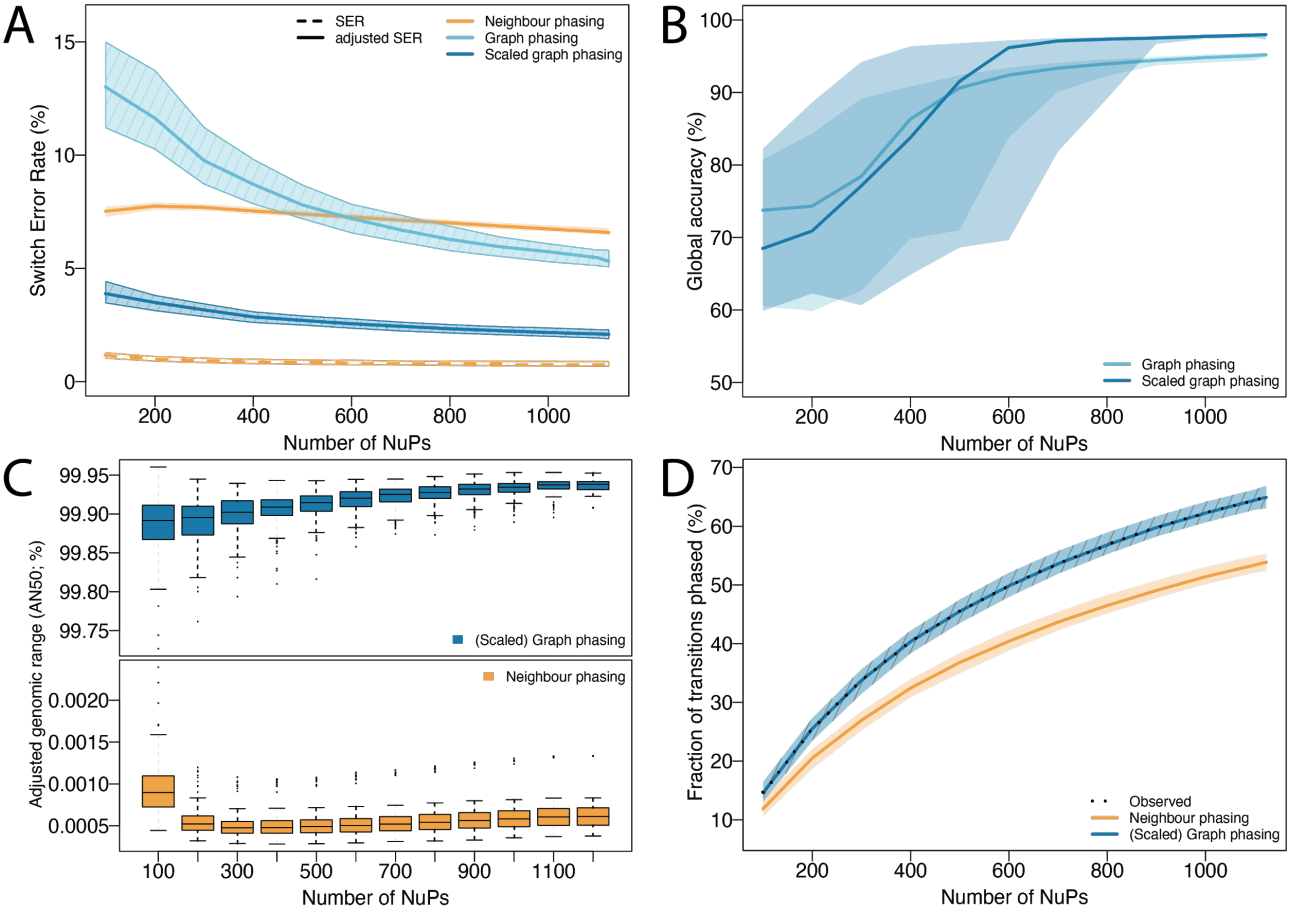

Figure 3: Quality of reconstructed haplotypes using the neighbour phasing algorithm (orange) and the basic (light blue) and proximity-scaled (blue) graph phasing algorithm. Haplotypes were predicted from an increasing number of included NuPs, with 10 iterations of random sampling of NuPs each. Lines show the median value, shaded areas indicate the interquartile range of results from 10 iterations of 19 mouse autosomes. **A) Switch Error Rate** (SER): Due to the large number of unconnected phased blocks, results from neighbour phasing show very low SER even with low sample size (dashed orange line). Adjusting the SER by introducing a penalty for unphased

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

transitions shows the impact of low completeness on neighbour phasing performance (solid orange line). Graph phasing performance improves with the number of NuPs considered and the proximity-scaled graph phasing shows lowest SER overall (dark blue line). Lines showing SER and adjusted SER overlap due to the small number of unconnected phasing blocks resulting from graph phasing. **B) Global Accuracy**: Graph phasing builds one main chromosome-spanning block which is susceptible to single local phasing errors. Results substantially improve with increasing sample size as more evidence for the correct haplotype structure is collected and proximity scaling improves performance overall. Neighbour phasing results are not shown due to the unknown relationship between the large number of unconnected blocks. **C) Phasing completeness (adjusted N50, AN50):** Graph phasing shows high completeness even for a low number of NuPs (dark blue), which is independent of proximity scaling. Almost all SNVs are phased within one major, chromosome-spanning haplotype block. Neighbour phasing yields a large number of small fragmented blocks (median 10 SNVs) with low overall completeness (orange). **D) Percent of phased transitions**. All pairwise transitions between neighbouring SNVs are considered. Between 14.5% and 62.2% of all transitions in the F123 genome were observed. Graph phasing phases 99.96% of all observed transitions, whereby neighbour phasing only phases 83.02% of all observed transitions.

Graph phasing performance. The additional higher-order phasing information considered by the graph phasing algorithm substantially improved the completeness of the reconstructed haplotypes independent of proximity scaling (Figure 3C). 99.97% (±0.004%) of input SNVs were phased into haplotype blocks (Figure 3D), 99.94% (±0.01%) of them into one main haplotype block (S50), spanning more than 99.99% (± 0.00003%) of the phasable genome (N50). Adjusting the span of the largest block by the fraction of phased SNVs yields an AN50 value of 99.94% (±0.010%) (Figure 3C). The graph phasing algorithm thus reconstructs dense chromosome-spanning haplotypes (Table 1).

Considering larger SNV windows increases the risk of integrating incorrect phasing information from co-observed SNV pairs located on homologous chromosome copies. Consequently, the accuracy of reconstructed haplotypes is lower than with strict neighbour phasing. The basic graph phasing approach yielded reconstructed haplotypes with a mean 95.14 % (±0.56%) and median 95.20% (IQR: 25.66%) global accuracy (Figure 3B) and 5.42% (±0.50%) switch errors (Figure 3A). To improve local accuracy while maintaining completeness we applied proximity scaling to the graph phasing approach. Proximity scaling increased global accuracy in general (median: 97.98%, IQR: 2.26%, Figure 3B), but the high standard deviation of ±8.45% at a mean of 94.29% indicates the presence of outliers. When a switch error occurs within a haplotype block, the assignment of subsequent alleles is inverted, reducing global accuracy while maintaining SER. The low occurrence of switch errors at a rate of 2.09% (±0.26%) demonstrates the improved performance compared to the basic graph phasing algorithm (Figure 3A). Thus proximity-scaled graph phasing shows best performance overall and results in accurate, chromosome-spanning haplotypes.

Table 1: Comparison of quality measures for the neighbour phasing algorithm, basic and proximity-scaled graph phasing algorithm for the full dataset. The mean of per-chromosome values is reported, standard deviation in brackets.

	Neighbour phasing	Graph phasing (basic)	Graph phasing (proximity-scaled)
% phased SNVs	95.94 % (±0.25)	99.97 % (±0.004)	
S50	10.84 SNVs (±0.5)	617,561.5 SNVs / 99.94 % (±0.010)	
N50	741.74 bp (±40.54)	126,454,374 bp / > 99.99 % (±0.00003)	

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AN50	-	126,374,367 bp / 99.94 % (±0.010)	
Global accuracy	99.00 % (±0.15)	95.14 % (±0.56)	94.29 % (±8.45)
SER	0.76 % (±0.13)	5.42 % (±0.50)	2.09 % (±0.26)

3.3.2 Performance at lower SNV density

The F123 mESC cell line has a relatively high SNV density (8 SNVs per 1kbp) compared to humans (approximately 1-1.5 SNVs per 1kbp, (1000 Genomes Project Consortium et al. 2015)). To show the effect of SNV density on the quality of haplotype reconstructions, we randomly subsampled the F123 SNV set to resemble human SNV density and evaluated the resulting haplotypes. In order to obtain an average SNV density of 1 SNV per 1kb, we retained 2,462,745 (13.57%) out of the known 18,150,228 F123 SNVs in the 2.46 billion bp mm10 mouse reference genome. The distribution of SNVs along the parental chromosomes remained constant (full SNV set: 87.11% CAST, 12.89% J129; subsampled: 87.14% CAST, 12.86% J129). Variants were randomly subsampled from the true parental haplotypes irrespective of their observation in the GAM NuPs. Similar to the full dataset (64.69% of known SNVs observed), 64.66% of all SNVs were observed in the subsampled dataset.

We explored accuracy and completeness of the best-performing proximity-scaled graph phasing algorithm on the subsampled dataset. All parameters, including the proximity scaling parameters, remained unchanged for the haplotype reconstruction. Despite the reduced SNV density and thus increased genomic distance between co-observed SNVs, GAMIBHEAR reconstructed accurate, dense, chromosome-spanning haplotypes: 99.96% of input SNVs were phased into haplotype blocks of minimum size 2, on average 99.95% (± 0.0096%) of those were phased in the main, chromosome-spanning haplotype block, covering 100% (± 0.00%) of the phasable genome. The mean global accuracy of 87.46% is still fairly high, the high standard deviation of ± 15.21% indicates a large span in the results. The median global accuracy of 96.64% and the switch error rate of 4.84% (± 0.6%) show that the quality of the reconstructed haplotypes in a subsampled dataset is only slightly different from that of the haplotypes reconstructed from the full dataset, indicating that the algorithmic approach is largely independent of SNV density and thus applicable to human data.

3.3.3 Time and Memory usage

Running GAMIBHEAR on the full 1123 NuP GAM dataset and phasing 11,741,055 heterozygous variants took on average 9.8 min elapsed time per chromosome using the neighbour phasing algorithm, 19.1 min using the basic and 20.4 min using the proximity-scaled graph phasing algorithm with a set window size of 20k SNVs on a desktop PC with 64GB of RAM without parallelisation. The neighbour and graph phasing algorithms required on average 7.6GB and 30GB per chromosome, respectively. Reducing or increasing the window size only marginally affected the performance of the methods in terms of completeness or accuracy; however, it did show a definite impact on the runtime and memory usage and changes to the default parameters should be made with care in order to assure successful completion of calculations. Reconstructing haplotypes from the dataset subsampled to human SNV density using the proximity-scaled graph phasing algorithm in sequential mode took 2.5 min on average per chromosome.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.4 Comparison with WhatsHap

We explored if the spatial phasing information from GAM data could be readily transformed for the use in existing algorithms. One approach is to combine the reads of captured genomic regions of single NuPs into long, chromosome-spanning pseudo reads to use with the long-read MEC solver WhatsHap (Patterson et al. 2015). This transformation could possibly deliver the phasing information captured with GAM experiments. WhatsHap tackles the NP-hard minimum error correction (MEC) problem with a fixed parameter tractable (FPT) approach, with its only parameter being the read coverage of SNVs. The algorithm is limited by a maximum coverage of 25 reads per SNV. Upon exceeding this threshold, the most informative reads are selected using a heuristic. Since internally WhatsHap uses a read by SNV matrix representation of genomic data, we directly transformed the GAM data for chromosome 1 into a 1105 pseudo reads by 974,770 heterozygous variant matrix. Within chromosome 1, 0.0074% of SNVs exceeded the upper coverage limit of 25 reads, forcing WhatsHap to select the most informative reads to meet the maximum read coverage constraint. That selection process led WhatsHap to discard the majority of reads (72 out of 1105 input reads remained), which consequently led to a loss of the majority of SNVs captured: only 14,400 SNVs (< 1.5% of all input SNVs) were kept. Out of these, 14,395 SNVs were phased and 5 SNVs showed equal evidence for both haplotypes. The reconstructed haplotype was not chromosome-spanning. The largest block contained approximately half of the phased variants (7367 SNVs, 51.16%). Accuracy was nonetheless high with 94.80% of considered variants phased correctly (compared to a global accuracy of 98.03% using GAMIBHEAR on chromosome 1) and a low switch error of 2.54% (GAMIBHEAR: 1.98%). However, due to the discard of over 98% of observed SNVs, the phasing results of WhatsHap on transformed GAM data are not practical.

4 Discussion

The phasing problem has been extensively studied, and different approaches have been proposed to solve it. These approaches are typically specific to and optimised for certain experimental designs and datatypes, such as Hi-C (Edge, Bafna, and Bansal 2017) and long reads (Patterson et al. 2015). Although both GAM and Hi-C capture the spatial proximity of SNVs in the nucleus, GAM data shows fundamentally different characteristics which makes existing methods ill-suited for GAM data. In Hi-C experiments, chimeric fragments are generated with parts originating from sequentially distal but spatially proximal functionally interacting genomic regions. When both parts of a chimeric fragment span at least one SNV each, the fragment provides long-range pairwise phasing information, because interactions between homologous chromosome pairs (h-trans errors) are captured rarely (Selvaraj et al. 2013)) although they do occur as frequently as intra-chromosomal interactions, but at further spatial distance and are thus not as efficiently captured in Hi-C experiments (Maass et al. 2018). In contrast, in GAM, while spatially close genomic regions are more likely to be captured within the same nuclear slice, GAM almost always captures fragments of both parental copies of a chromosome. Hence, the combination of all individual SNV-spanning reads of a NuP provides potential phasing information, albeit locally constrained. This property also separates GAM NuPs from long reads, where multiple SNVs can be captured by a single read and these SNVs are then guaranteed to have originated from the same parental chromosomal copy. To ascertain these differences we tested GAM data on the long-read MEC solver WhatsHap. While initial results were not convincing, we believe the results could be improved, for example by breaking the NuP into multiple reads of consecutive captured genomic regions using a distance threshold between captured SNVs. With GAMIBHEAR we implemented a phasing strategy based on GAM data characteristics and directly applicable to GAM data in the form of the presented neighbour and

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

graph phasing approaches. We did not attempt to transform GAM data for use with HapCut2, as it has been well known and stated by the authors that the performance of HapCut2 strongly depends on the correct error model being used and no such model exists for GAM data (Edge, Bafna, and Bansal 2017).

In comparison to Selvaraj et al. (2013), who reconstructed F123 haplotypes using HaploSeq, a phasing approach which combines proximity ligation experiments with the HapCUT algorithm, the chromosome-spanning largest blocks resulting from GAMIBHEAR and HaploSeq both span over 99.99% of the phasable genome. However, the major block from GAMIBHEAR's proximity-scaled graph-phasing algorithm includes >99.9% of observed variants compared to about 95% of observed variants using HaploSeq, an improvement due to the large genomic span covered by GAM NuPs. Although the graph phasing algorithm generates highly complete results from its input data even at low coverage, a drawback of the presented method is in the sparsity of the data itself. While in the Hi-C data of Selvaraj et al. (2013) 99.6% of variants were covered by at least one read, in the GAM data set only 64.69 % of variants are captured. While this creates no challenge in the generation and analysis of highly accurate and informative 3D chromatin contact maps from GAM data, it does affect the overall completeness of reconstructed haplotypes. Co-phasing of sequentially close by but uncovered SNVs or incorporation of statistical phasing provide means of expanding the reconstructed haplotypes by uncovered SNVs.

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 this probabilistic model we observed a stark decline in phasing information within 10 Mb distance from the source SNV. This decline is likely due to the formation of highly interacting genomic regions and corresponding organisational chromatin structures such as self-interacting TADs (Mb 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 MaxST obtained 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 SNVs 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 analysis, no phasing blocks were broken due to proximity scaling of edge weights.
Relevant allele-specific research of single genes or small genomic regions describes the primary use case of the highly accurate neighbour-phasing. This approach is most suitable if the major interest concentrates on local phasing results such as when disease causing genes with disturbed expression are in focus and no chromosome-spanning haplotypes are of need.

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. Application of our proximity-scaled graph phasing algorithm on F123 GAM data downsampled to human SNV density suggests that the reconstruction of haplotypes is suitable and well applicable for the use in human data as well.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5 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 open questions in this field and thus 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.

6 Author contributions

JM performed bioinformatic analysis on GAM data and implemented the algorithms and R package. JM, BK, SR and RFS designed the algorithms. RK, GL and AK 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.

7 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). SR is supported by DFG Collaborative Research Center SFB 876, subproject C1. JM is supported by DFG Priority Program SPP2202 “Spatial Genome Architecture in Development and Disease”.

8 References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

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

Bafna, Vineet, Sorin Istrail, Giuseppe Lancia, and Romeo Rizzi. 2005. “Polynomial and APX-Hard Cases of the Individual Haplotyping Problem.” *Theoretical Computer Science* 335 (1): 109–25.

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.

Beagrie, Robert A., Christoph J. Thieme, Carlo Annunziatella, Catherine Baugher, Yingnan Zhang, Markus Schueler, Dorothee C. A. Kramer, et al. 2020. “Multiplex-GAM: Genome-Wide Identification of Chromatin Contacts Yields Insights Not Captured by Hi-C.” https://doi.org/10.1101/2020.07.31.230284.

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*

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Chandradoss, Keerthivasan Raanin, Prashanth Kumar Guthikonda, Srinivas Kethavath, Monika Dass, Harpreet Singh, Rakhee Nayak, Sreenivasulu Kurukuti, and Kuljeet Singh Sandhu. 2020. "Biased Visibility in Hi-C Datasets Marks Dynamically Regulated Condensed and Decondensed Chromatin States Genome-Wide." *BMC Genomics* 21 (1): 175.

Chen, Haiming, Sijia Liu, Laura Seaman, Cyrus Najarian, Weisheng Wu, Mats Ljungman, Gerald Higgins, Alfred Hero, Max Wicha, and Indika Rajapakse. 2017. "Parental Allele-Specific Genome Architecture and Transcription during the Cell Cycle." https://doi.org/10.1101/201715.

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, Thomas Huebsch, Gayle McEwen, Eun-Kyung Suk, and Margret R. Hoehe. 2010. "ReFHap: A Reliable and Fast Algorithm for Single Individual Haplotyping." In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, 160–69. BCB '10. New York, NY, USA: Association for Computing Machinery.

Edge, Peter, Vineet Bafna, and Vikas Bansal. 2017. "HapCUT2: Robust and Accurate Haplotype Assembly for Diverse Sequencing Technologies." *Genome Research* 27 (5): 801–12.

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.

Fiorillo, Luca, Francesco Musella, Rieke Kempfer, Andrea M. Chiariello, Simona Bianco, Alexander Kukalev, Ibai Irastorza-Azcarate, et al. 2020. "Comparison of the Hi-C, GAM and SPRITE Methods by Use of Polymer Models of Chromatin." https://doi.org/10.1101/2020.04.24.059915.

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.

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.

Hu, T. C. 1961. "Letter to the Editor—The Maximum Capacity Route Problem." *Operations Research* 9 (6): 898–900.

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.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.
Lo, Christine, Ali Bashir, Vikas Bansal, and Vineet Bafna. 2011. "Strobe Sequence Design for Haplotype Assembly." *BMC Bioinformatics* 12 Suppl 1 (February): S24.
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.
Maass, Philipp G., A. Rasim Barutcu, Catherine L. Weiner, and John L. Rinn. 2018. "Inter-Chromosomal Contact Properties in Live-Cell Imaging and in Hi-C." *Molecular Cell* 69 (6): 1039–45.e3.
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.
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.
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.
Rivera-Mulia, Juan Carlos, Andrew Dimond, Daniel Vera, Claudia Trevilla-Garcia, Takayo Sasaki, Jared Zimmerman, Catherine Dupont, Joost Gribnau, Peter Fraser, and David M. Gilbert. 2018. "Allele-Specific Control of Replication Timing and Genome Organization during Development." *Genome Research* 28 (6): 800–811.
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.
Tourdot, Richard W., and Cheng-Zhong Zhang. 2019. "Complete Haplotype Determination and Single-Chromosome Analysis." *bioRxiv*. https://doi.org/10.1101/629337.
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.
Zahn, Laura M. 2020. "Effects of Allele-Specific Open Chromatin." *Science* 369 (6503): 519–21.

Supplement

S1 Benchmark Genome

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). With the haplotype structure thus known, this cell line serves as the benchmark for all downstream experiments and analyses.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927061; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

number SRX037820), respectively. To determine the haplotypes of the F123 line, WGS reads were trimmed using Cutadapt (Martin 2011) and mapped to the mouse reference genome mm10 using BWA (Heng Li and Durbin 2009). SNVs were identified using bcftools (Heng Li 2011) and SNVs covered by <5 reads and quality <30 were excluded.

S2 GAM dataset, pre-processing and quality control

1261 individual GAM NuPs of the F123 line were obtained from the 4D Nucleome Consortium data portal under accession number 4DNBSTO156AZ. The F123 SNVs 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.

For quality assessment of each sample, the genome was split into fixed windows of size 50kb. For each NuP 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.of all windows for NuPi 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 NuPs then determined the sample-specific nucleotide coverage thresholds t_i (in bp) for each NuP. 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.
NuPs selected for further analysis had < 60% orphan windows and > 20,000 uniquely mapped reads. 1123 NuPs (89%) passed these quality thresholds.
Reads were then counted at known heterozygous SNV positions using samtools mpileup (H. Li et al. 2009). Because of the frequently low coverage from independent (i.e. non-duplicate) reads at most positions (30% of observed SNVs are covered by 2 or less reads, 50% by 5 or less reads), we counted an allele as present if it was observed in at least one read at the examined position.