EVALUATING AND CREATING GENOMIC TOOLS FOR CASSAVA BREEDING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Ariel Chan

May 2019

EVALUATING AND CREATING GENOMIC TOOLS FOR CASSAVA BREEDING

Ariel Chan, Ph. D.

Cornell University 2019

The genetic improvement of *Manihot esculenta*, or cassava, has historically been slow, largely because its biology renders traditional breeding techniques inefficient and because of little interest from the private sector. The goal of the Next Generation Cassava Breeding project (NEXTGEN) is to assist breeding institutions in Nigeria, Uganda, and Tanzania with increasing the rate of genetic improvement of cassava through implementation of genomic selection (GS). The three chapters of my thesis outline my work and involvement with the NEXTGEN project. The first chapter details our investigation of two questions: 1) can we use existing imputation methods developed by the human genetics community to impute missing genotypes in datasets derived from non-human species and 2) are these methods, which were developed and optimized to impute ascertained variants, amenable for imputation of missing genotypes at next-generation sequencing (NGS)-derived variants? In the second chapter, we introduce a statistical method, BIGRED (Bayes Inferred Genotype Replicate Error Detector), for detecting mislabeled and contaminated samples using shallow-depth sequence data. BIGRED addresses key limitations of existing approaches and produced highly accurate results in simulation experiments. In the third chapter, we outline how we used the multi-generational pedigree and genotyping-by-sequencing (GBS) data from the International Institute

of Tropical Agriculture (IITA) to characterize the recombination landscape across the 18 chromosomes of cassava. We detected SNP intervals containing crossover events using SHAPEIT2 and duoHMM, constructed a genetic map using these intervals, compared it to an existing map constructed by the International Cassava Genetic Map Consortium (ICGMC), and constructed sex-specific genetic maps to see if cassava displays sexual dimorphism in crossover distribution and frequency.

BIOGRAPHICAL SKETCH

Ariel Wan-Yee Chan completed a Bachelor's of Science in Agriculture at the University of Georgia, majoring in Food Science and Technology. During her last semester as an undergraduate, she studied abroad at the University of Natural Resources and Life Sciences in Vienna, Austria. Heeding the advice of her older sister, Ariel applied for a National Science Foundation Graduate Research Fellowship in 2012. She received the fellowship and began graduate school at Cornell University in August 2012. In May 2019, Ariel completed her graduate degree in Plant Breeding and Genetics, with a minor in Computation Biology and International Agriculture under the supervision of Jean-Luc Jannink, Martha Hamblin, Amy Williams, and Ronnie Coffman.

ACKNOWLEDGEMENTS

I would like to thank my parents Cheong-Wo Hunter Chan and Wing-Wah Ruby Chan for giving me a life of privilege and my older sister Leslie Chan for drafting a blueprint for success in academia. Hard work alone would not have been sufficient to get to and through graduate school. I would like to thank two professors who take their roles as educators very seriously: Jason Mezey and John Tsitsiklis. Jason Mezey, a professor at Cornell University, taught me quantitative genomics and genetics. He gave me the language and vocabulary required to navigate the field of statistical genomics and helped me get a footing during the first years of graduate school. John Tsitsiklis, a professor at the Massachusetts Institute of Technology, taught me probability theory and statistical inference through MIT OpenCourseWare. He made his thought process completely transparent, showing me how he logically tackles a problem. I would like to thank Amy Williams, Jean-Luc Jannink, and Ronnie Coffman for serving on my committee and providing helpful feedback. I would like to thank the National Science Foundation for awarding me a graduate research fellowship. Lastly, I would like to thank Thomas Ehrmann, my life partner, and San Mao, my feline friend, for providing moral support and companionship through these years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

EVALUATING IMPUTATION ALGORITHMS FOR LOW-DEPTH GENOTYPING-BY-SEQUENCING (GBS) DATA[1]

ABSTRACT

Well-powered genomic studies require genome-wide marker coverage across many individuals. For non-model species with few genomic resources, high-throughput sequencing (HTS) methods, such as Genotyping-By-Sequencing (GBS), offer an inexpensive alternative to array-based genotyping. Although affordable, datasets derived from HTS methods suffer from sequencing error, alignment errors, and missing data, all of which introduce noise and uncertainty to variant discovery and genotype calling. Under such circumstances, meaningful analysis of the data is difficult. Our primary interest lies in the issue of how one can accurately infer or impute missing genotypes in HTS-derived datasets. Many of the existing genotype imputation algorithms and software packages were primarily developed by and optimized for the human genetics community, a field where a complete and accurate reference genome has been constructed and SNP arrays have, in large part, been the common genotyping platform. We set out to answer two questions: 1) can we use existing imputation methods developed by the human genetics community to impute missing genotypes in datasets derived from non-human species and 2) are these methods, which were developed and optimized to impute ascertained

[1] **A. W. Chan**, M. T. Hamblin, and J.-L. Jannink, "Evaluating Imputation Algorithms for Low-Depth Genotyping-By-Sequencing (GBS) Data.," *PLoS One*, vol. 11, no. 8, p. e0160733, 2016.

variants, amenable for imputation of missing genotypes at HTS-derived variants? We selected Beagle v.4, a widely used algorithm within the human genetics community with reportedly high accuracy, to serve as our imputation contender. We performed a series of cross-validation experiments, using GBS data collected from the species *Manihot esculenta* by the Next Generation (NEXTGEN) Cassava Breeding Project. NEXTGEN currently imputes missing genotypes in their datasets using a LASSO-penalized, linear regression method (denoted 'glmnet'). We selected glmnet to serve as a benchmark imputation method for this reason. We obtained estimates of imputation accuracy by masking a subset of observed genotypes, imputing, and calculating the sample Pearson correlation between observed and imputed genotype dosages at the site and individual level; computation time served as a second metric for comparison. We then set out to examine factors affecting imputation accuracy, such as levels of missing data, read depth, minor allele frequency (MAF), and reference panel composition.

INTRODUCTION

Well-powered genomic studies require genome-wide marker coverage across many individuals. Many genotyping methods exist, and one typically selects a genotyping platform based on budgetary constraints and the available molecular tools for the species in question. Genetic variation in the human genome, for instance, has largely been captured using single-nucleotide polymorphism (SNP) arrays that can assay up to 2.5 million variants [1]. The per-sample and array-design costs of these assays, however, make them accessible only to well-funded model

systems. For species lacking a complete reference genome or predesigned high-density SNP genotyping arrays, high-throughput sequencing (HTS) methods, such as Genotyping-By-Sequencing (GBS), offer an economic approach for surveying variants at the genome level. The multiplex capabilities of HTS methods allow for great flexibility in experimental design. For instance, given a fixed number of sequencing reads and genome size, one can choose to sequence a small number of individuals, allocating the reads among a small number of individuals, or one can choose to distribute the reads among a larger sample of individuals. The former framework generates datasets with relatively low levels of missing data. The small sample size limits the number of detected variants, but this may be a moot point depending on the biological question one wishes to address. For studies requiring large sample sizes and dense genome-wide marker coverage, e.g. genome-wide association studies (GWAS) and genomic selection (GS), the latter genotyping framework is preferable, and one can impute or infer missing genotypes with appropriate imputation methods [2].

Genotype imputation is a well-established statistical technique for estimating unobserved genotypes. Many genotype imputation algorithms and software packages exist, but most were primarily developed by and optimized for the human genetics community, a field where a complete and accurate reference genome has been constructed and SNP arrays have, in large part, been the common genotyping platform. These algorithms differ in their details but all essentially pool information across individuals in either a study sample or a reference panel or both to estimate haplotype frequencies from the observed genotype data, imputing missing

genotypes simultaneously. Although the statistical methods for genotype imputation are now highly developed and widely used, selecting the set of haplotypes to include in the reference panel for maximum imputation accuracy in a given study population remains unclear. Selection schemes typically take one of two approaches: a 'best match' approach, which attempts to construct a reference panel that closely matches the ancestry of the study sample, or a 'cosmopolitan' approach, which makes use of all available haplotypes [3].

To assess the applicability of human-tailored imputation algorithms in non-model species datasets, we evaluated the imputation performance of Beagle v.4, a widely used haplotype-phasing algorithm with reportedly high accuracy, in low-depth GBS-generated data collected from the species *Manihot esculenta* (commonly referred to by its colloquial name 'cassava'). We compared Beagle v.4 to a LASSO-penalized, linear regression imputation method (denoted glmnet). We chose Beagle v.4 over other haplotype-phasing programs because the algorithm 1) scales well to large sample sizes (>1000) while other algorithms require some form of parameter space reduction to be computationally competitive, 2) requires no parameter specification, e.g. effective population size, 3) takes genotype likelihoods as input, and 4) performs genotype calling [4]. The Next Generation (NEXTGEN) Cassava Breeding Project currently employs glmnet to impute missing genotypes in NEXTGEN datasets; we selected glmnet to serve as a benchmark method for this reason. Glmnet takes a linear regression approach to genotype imputation. The algorithm assumes that any locus on a given chromosome can be modeled as a linear combination of other intra-chromosomal loci, independent of locus distance

and locus order. Such methods model only the statistical correlations between loci and make no attempts at relating observed correlations to underlying biological phenomena, such as linkage disequilibrium (LD; the nonrandom association of alleles among linked loci). Results from [5] show that imputation of unordered markers can be accurate, particularly when LD between markers is high and when individuals in the study sample share recent common ancestry.

We evaluated Beagle and glmnet under three imputation scenarios: imputation guided by 1) no reference panel, 2) a reference panel with large genetic diversity (reference panel A), and 3) a reference panel that closely matches the ancestry of the study sample (reference panel B). We describe the composition of reference panel A and B in greater detail in the Methods and Materials section. We provide a schematic drawing of reference panel A and B in Appendix Figure 1.1A and of the three imputation scenarios in Appendix Figure 1.1B. We performed a series of cross-validation experiments using GBS data collected from the species *Manihot esculenta* by NEXTGEN. For simplicity, we focused on the situation where the reference haplotypes in scenario 2 and 3 are defined on the same set of polymorphic sites as those found in the study sample. For each cross-validation experiment, we measured imputation accuracy at both the site- and individual-level, using the sample Pearson correlation statistic as an estimate of accuracy. We assessed the impact of missing data, read depth, minor allele frequency (MAF), and reference panel composition on imputation accuracy. We report the computation requirement and a scalar summary of imputation accuracy measured at the site and individual for Beagle and glmnet under each scenario.

MATERIALS AND METHODS

We evaluated the performance of Beagle and glmnet under three imputation scenarios using data collected at biallelic SNPs on chromosome 5 from two NEXTGEN cassava populations: the International Institute of Tropical Agriculture's (IITA) Genetic Gain (GG) population, a collection of historically important clones, and IITA's Cycle 1 (C1) population. We first describe how the sequence data was generated and processed then provide a description of the two IITA populations.

Data generation and variant calling

*Ape*KI GBS libraries were constructed at the Institute for Genomic Diversity at Cornell University and sequenced on the Illumina HiSeq 2000/2500 at the Biotechnology Resource Center at Cornell University following the protocol outlined in [6]. Converting the raw read data into a final set of SNP calls involved a number of steps; a complete description of the protocol is beyond the scope of this paper. We refer the reader to [7] and https://bitbucket.org/tasseladmin/tassel-5-source/wiki/Tassel5GBSv2Pipeline for a detailed description of version 4 and 5 of the TASSEL-GBS bioinformatics pipeline, respectively. SNPs were extracted from the raw sequence data using the TASSEL 5.0 GBS discovery pipeline with alignment to the *Manihot esculenta* v.6 assembly. Sequence reads generated by GBS assays were trimmed or padded to 64 bases and subjected to quality filters (refer to section 'Favoring allelic redundancy over quality scores' of [7]). The filtered sequence reads were aligned to the cassava reference genome version 6 assembly. Genotype calling

then proceeded for each individual by counting the number of times each allele was observed and using empirically determined thresholds for genotype calls. SNP calling was then performed using the inferred genotypes. To minimize ascertainment bias, all NEXTGEN samples (in addition to non NEXTGEN samples) sequenced to date were used for variant detection. Putative SNPs were filtered based on a minimum minor allele frequency (mnMAF) of 0.001. NEXTGEN opted to use a relatively low-stringency filter since false-positive variants can be filtered out in subsequent steps. We obtained 18 VCF files (one VCF file per chromosome) after processing the raw GBS sequence reads from NEXTGEN samples. The raw VCF file for chromosome 5, a chromosome approximately 30 Mbp in length, contained 30018 entries (variant sites) and 15750 samples, 164 of which were blank negative controls. Appendix Figure 1.2 shows the distribution of variants across the length of chromosome 5. As of writing this manuscript, the data we analyzed are free and publically available at www.cassavabase.org.

Each sample ID (i.e. column name) in the VCF files follows the following format: 'ShortName:LibraryPrepID'. Upon closer examination, we found 554 'ShortNames' that appear >2 times in the VCF file for chromosome 5. Samples sharing an identical 'ShortName' represent (supposed) technical or biological replicates of a unique individual. Before merging the sequence data from samples sharing an identical 'ShortName', we applied an Expectation-Maximization (EM) algorithm to detect mislabeling of samples among technical and biological replicates (unpublished). We merged the sequence data for cases where the algorithm detected no error. We then removed non-biallelic sites from the dataset, leaving a

total of 20302 biallelic SNPs for analysis. Appendix Figure 1.2 shows the distribution of biallelic SNPs across the length of chromosome 5.

The FORMAT field of the VCF file consists of five colon-separated, sub-fields: genotype (GT), allelic read depth (AD), read depth (DP), genotype quality (GQ), and Phred-scaled likelihood (PL). For our purposes, we were interested in only the AD subfield, which encodes the observed counts of each of the two alleles in individual d at site v: $X_d^{(v)} = \left(N_A^{(v,d)}, N_B^{(v,d)}\right)$, where $N_A^{(v,d)}$ and $N_B^{(v,d)}$ denote the observed counts of allele A and allele B, respectively, in individual d at site v. To ensure that genotype likelihoods were calculated in a consistent manner, we computed genotype likelihoods for each individual at each site using the data stored in the AD subfield rather than using those provided in the PL subfield of the VCF file. Given observed data $X_d^{(v)}$ and fixed sequencing error rate e = 0.01, we computed the likelihood for genotype $G_d^{(v)} = g$. We calculated genotype likelihoods for a single individual at a single site independent of all other individuals and sites in the sample using the following equation:

$$P\left(X_d^{(v)} | G_d^{(v)} = g, e\right) = \binom{N_A^{(v,d)} + N_B^{(v,d)}}{N_B^{(v,d)}} (1 - p_B)^{N_A^{(v,d)}} (p_B)^{N_B^{(v,d)}}$$

$$p_B = \begin{cases} e, & \text{when } g = AA \\ 0.50, & \text{when } g = AB \\ 1 - e, & \text{when } g = BB \end{cases}.$$

We estimated posterior probabilities for the three genotypes using the likelihoods defined above and assuming a uniform genotype prior. We summarized posterior probabilities into genotype dosages since the glmnet algorithm can only

take scalar-valued genotypes as input. Genotype dosages take values in [0,2] or NA for the case where no data is observed for a given individual at a site. We converted genotype likelihoods into normalized, Phred-scaled likelihoods to use as input for Beagle.

Germplasm

IITA has a large GG population for which there are many years of historical phenotype data collected in many environments. NEXTGEN selected a subset of GG individuals to serve as a training population (TP) for genomic selection (GS) at IITA. NEXTGEN selected an individual if plant material still existed for the individual (i.e. DNA could be extracted to obtain genotype data) and if phenotype records for the individual were based on a sufficient number of observations. As of writing this report, 694 individuals met these criteria [8]. From this point forward, we refer to these 694 individuals as the GG population. Genomic estimated breeding values (GEBVs) were obtained using the genomic best linear unbiased prediction (BLUP) method and the top GG individuals were selected to serve as founders of the IITA GS breeding program. To avoid inbreeding depression, NEXTGEN designed a crossing framework based on results from a k-means clustering analysis, crossing two GG individuals only if they belonged to different clusters. Based on pedigree records, a total of $y \geq 474$ crosses were made, with only a subset of these crosses (134 crosses using 82 individuals) producing viable progeny. The large variation in viable progeny number among attempted crosses results from the wide variation in flowering time, rate, and fertility in cassava [9]. Viable progeny from GG crosses

collectively form the C1 population. Two randomly sampled individuals from the C1 population are nominally related in one of three possible ways: the two individuals are 1) full siblings, 2) half siblings, or 3) unrelated. We have pedigree records for 2207 C1 individuals but found 2490 individuals in the VCF file whose sample IDs indicate C1 population membership (i.e. samples with sample name prefix "2013_" and "TMS13"). We used all 2490 C1 individuals as the target of imputation for scenarios 2 and 3.

Inconsistencies among sources of information (i.e. the pedigree record, the sequence data in the VCF file, and the list of 694 GG individuals) influenced the design of the two reference panels used in imputation scenarios 2 and 3. According to the pedigree record, 82 individuals gave rise to the C1 population; however, only 78 of these 82 supposed C1 parents appear in the list of 694 GG individuals. We expected all C1 parents to appear in the list of GG individuals. We found sequence data for these 78 individuals in the VCF file. Of the remaining four individuals listed as C1 parents in the pedigree record, we found sequence data for only two individuals B9200061 and B9200068 in the VCF file. We expected all C1 parents to have sequence data since this information was required for estimation of breeding values. We found no sequence data for individuals I970466 and I974769 in the VCF file.

The 694 GG individuals served as the reference panel for scenario 2 (reference panel A, representing a "cosmopolitan" reference panel). The 80 individuals listed as C1 parents in the pedigree record for whom we have sequence data served as the reference panel for scenario 3 (reference panel B, representing a

"best-match" reference panel). The intersection of reference panel A and panel B consists of 78 C1 parents. We provide a schematic drawing of reference panel A and B in Appendix Figure 1.1A.

The two reference panels collectively contain 696 unique individuals (the union of reference panel A and panel B). We performed a principal component analysis (PCA) to explore whether there is any evidence of population structure among the 696 reference panel individuals. We calculated the realized additive relationship matrix for the 696 reference panel individuals at a subset of the 20205 biallelic SNPs using the function "A.mat" from the R package "rrBLUP" [10], [11]. We excluded sites with >50% missing data (max.missing=0.5) from the calculation and imputed missing dosage values using the "EM" option (impute.method="EM"). We then performed PCA through eigenvalue decomposition of the realized additive relationship matrix (covariance matrix) using the R function "prcomp" and plotted the first two principal components (Appendix Figure 1.3). We observed little evidence of subpopulation structure among the 696 reference panel individuals.

Dataset for scenario 1 (imputation using no reference)

If each individual and each site in the study sample have a low proportion of missing data, no reference panel is needed to impute the missing genotypes in the sample; the almost complete data from the other individuals and the high marker density should provide sufficient information to impute with high accuracy. We tested this concept using the 694 GG individuals as our study sample. We extracted the genotype dosages and normalized, Phred-scaled likelihoods for the GG

individuals at biallelic sites (*n* = 20302). Appendix Figure 1.4 shows the distribution of the proportion of missing data per site. The term "missing" denotes zero reads observed at a given site for a given individual. We removed sites with >90% missing data, leaving a total of 20205 sites for cross-validation experiment 1. We use this same set of sites for imputation scenario 2 and 3 for reasons given in the proceeding section. Appendix Figure 1.5A and S5B show the distribution of the mean read depth per site averaged across all 694 GG individuals and across all 696 reference panel individuals, respectively.

Datasets for scenario 2 and 3

We assessed the impact of reference panel composition on imputation accuracy using C1 individuals (*n* = 2490) as the target of imputation. We constructed two reference panels, one designed to represent a cosmopolitan reference panel for imputation scenario 2 and the other designed to represent a best-match reference panel for scenario 3. Variants absent from the reference panel, but present in the study sample, cannot be imputed. We, therefore, focused on the situation where the reference panel is defined on the same set of polymorphic sites as those found in the study sample, using the same set of 20205 biallelic SNPs defined in scenario 1.

We extracted genotype dosages and normalized, Phred-scaled likelihoods for the 2490 C1 individuals. To construct the reference panels for scenario 2 and 3, which collectively consist of 696 individuals, we extracted genotype dosages and normalized, Phred-scaled likelihoods for the 696 reference panel individuals. We

ran the glmnet and Beagle imputation algorithms, using the extracted genotype dosages and normalize, Phred-scaled likelihoods for the 696 individuals as input, respectively. We constructed the cosmopolitan reference panel for Beagle (glmnet) using the inferred haplotypes (imputed genotype dosages) from the 694 GG individuals; we constructed the best-match reference panel for Beagle (glmnet) using the inferred haplotypes (imputed genotype dosages) from the 80 C1 parents. Although a reference panel cannot be explicitly specified when imputing with glmnet, the algorithm can still make use of the information encoded in non-study sample individuals. The increased sample size of the training data should, in theory, increase imputation accuracy.

Glmnet Algorithm

We used the R package glmnet to fit a LASSO-penalized, linear regression model to the observed genotype data [12]. The glmnet imputation algorithm described here employs a combination of both variable selection and the least absolute angle and selection operator (LASSO). LASSO penalized estimates are solutions to an optimization problem of the form:

$$\tilde{\beta} = argmin_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|^q \right\}.$$

We set q = 1. The variable λ is a regularization parameter that controls the trade-offs between lack of fit and model complexity; $\lambda \geq 0$ [13]. In addition to shrinking estimates toward zero, LASSO can perform variable selection, setting a subset of regression coefficients to zero [13]. The algorithm initializes by imputing missing genotypes at site *v* to the mean genotype at site *v*. Although the LASSO performs

variable selection on its own, we performed an initial round of variable selection to decrease computation time -- shrinking the variable search space to a subset of 60 markers rather than using all markers on a chromosome as potential predictors of genotype. We calculated pairwise correlations between the target marker and all intra-chromosomal markers, retaining the 60 markers that showed the strongest correlation with the target marker. We selected a maximum retention number of 60 arbitrarily. Other approaches for shrinking the variable search space exist but were not explored in this study. By default, glmnet selects a lambda value using 10-fold cross-validation, looking at 100 different lambda penalty coefficients. To decrease computation time, 5-fold cross validation was performed on 10 lambda values.

Beagle v.4

Beagle v.4 is an iterative algorithm for fitting a local haplotype hidden Markov model (HMM) to genotype data. The algorithm alternates between model building and sampling, using stochastic expectation maximization (EM) to converge towards the most probable solutions [14]. There are five components to an HMM: 1) hidden states, 2) observed values, 3) state-transition probabilities, 4) emission probabilities, and 5) initial-state probabilities [15]. The underlying hidden states of an HMM generate the observed data, and the state-transition probabilities, emission probabilities, and initial-state probabilities are parameters of the HMM. In the context of haplotype phase and missing genotype inference, the observed data are the unphased genotypes, while the hidden states represent haplotype membership

and the true, underlying genotypes. Beagle estimates state-transition probabilities, emission probabilities, and initial-state probabilities from the data.

The algorithm begins by imputing missing genotypes according to allele frequencies and randomly phasing heterozygous genotypes. Beagle v.4 then uses these initial haplotype estimates to obtain estimates of the HMM parameters. The algorithm constructs a directed acyclic graph (DAG) using the haplotype data and estimates the HMM parameters using observed haplotype counts and the assumption of Hardy-Weinberg Equilibrium (HWE). [16]. Browning provides a detailed explanation of how the algorithm constructs the graphical model in [16]. After constructing the model, Beagle samples four pairs of haplotypes per individual from the posterior distribution of haplotypes conditioned on the observed genotypes. These sampled haplotypes serve as input for the next iteration to re-estimate the model parameters. The model building and sampling procedure repeats for five burn-in iterations, followed by an additional five iterations. Beagle v.4 outputs a consensus haplotype for each individual, which is constructed from the 20 haplotypes sampled during the non burn-in iterations. In addition to consensus haplotypes, Beagle v.4 outputs imputed genotype dosages (also known as posterior mean genotypes) for each individual at each site. A reference panel can be specified in Beagle v.4 with the *ref* parameter. All genotypes in the reference panel must be non-missing and phased.

Measuring imputation accuracy

There are various metrics of imputation accuracy: imputation correlation, the Pearson correlation between observed and imputed genotypes, imputation concordance, the proportion of correctly imputed genotypes, imputation quality score (IQS), the concordance adjusted for chance agreement), etc. [17]. We selected the Pearson correlation coefficient to serve as our metric of imputation accuracy at the site level since its interpretation does not depend on MAF. The sample Pearson correlation between two variables is defined as the covariance of the two variables divided by the product of their standard deviations: $r = \frac{\sum_{i=1}^{L}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{L}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{L}(y_i-\bar{y})^2}}$.

When computing the sample Pearson correlation, *r*, at site *v*, *X* denotes the site's vector of observed genotype dosages and *Y* denotes the site's vector of imputed genotype dosages. The sample Pearson correlation is calculated with the assumption that the genotype dosages are accurately estimated. The sample Pearson correlation is a function of two vectors, both of length *L*. The value of *L* varies across sites for two reasons: the random nature of the masking scheme and non-uniform representation of sites within the set of validation genotypes defined by Caller A and B. The Pearson correlation coefficient is undefined when either $L<2$ or when the vector of imputed genotype dosages is invariant.

To calculate imputation accuracy, we masked a set of validation genotype dosages, imputed, and calculated the sample Pearson correlation between observed and imputed genotype dosages. We employed two different methods, Caller A and Caller B, to define the set of validation genotypes for cross-validation experiments. Caller A returns a genotype dosage for individual *d* at site *v* if individual *d* was

surveyed a minimum of seven times at site *v* and returns NA otherwise. The second method, Caller B, returns a genotype dosage for individual *d* at site *v* if the most likely genotype is at least 10 times more likely than the second most likely genotype and returns NA otherwise. We found that cross-validation experiments using Caller A and B validation genotypes returned similar results for imputation scenario 1 (data not shown), resulting in our decision to run scenario 2 and 3 using only Caller B validation genotypes.

We simulated a scenario where genotypes were missing in a random fashion across the genome and obtained estimates of imputation accuracy using 10-fold cross validation. The masking scheme is best visualized by describing the datasets as matrices, where the rows represent biallelic sites and the columns represent individuals. The elements in a matrix represent genotypes: individual *d* has genotype $G_d^{(v)} = g$ at marker *v*. We extracted each genotype's read depth from the VCF file using VCFtools [18]. We partitioned the set of validation genotypes into 10 equally sized, disjoint subsets: *M1*, *M2*, …, *M10*. Each subset corresponds to a fold in the 10-fold cross-validation scheme. As an example, we generated the masked dataset for fold 1 by taking the original data matrix, finding the coordinates of the genotypes belonging to the set *M1*, and setting the elements in these coordinates to missing. This masking scheme resulted in 10 masked datasets (i.e. 10 folds). We calculated the imputation accuracy on a per-site basis for each fold and the imputation accuracy on a per-individual basis for each fold. We then calculated the median imputation accuracy per-marker and the median imputation accuracy per-individual across the 10 folds.

Measuring computation cost

We measured computation time as the number of CPU minutes required to complete the imputation of one dataset. All jobs were submitted to the Computational Biology Service Unit at Cornell University, which uses an eight core Linux (Centos 6.2) Dell PowerEdge M600 with 16GB RAM.

RESULTS

Imputation with No Reference Panel

We imputed masked genotypes at 20205 SNPs on chromosome 5 in a sample of 694 individuals from the GG population. In this section, we report the results from cross-validation experiments where the set of validation genotype dosages was defined using Caller B (see Methods).

The sample Pearson correlation is a function of two vectors, both of length L. The value of L varies across sites and individuals because genotype masking occurs at random and because genotype call rates vary across the 20205 sites (see Methods). The sample correlation coefficient at site v is undefined under two scenarios: when $L<2$ (true for 34 of the 20205 sites in the dataset) and when the vector of imputed genotype dosages at site v has a variance equal to zero. The latter occurs when imputation returns identical genotype dosages for all L masked genotypes at site v. We obtained accuracy estimates for Beagle at 13028 sites (set A) and 19933 sites (set B) for glmnet. Set A is a subset of B, i.e. every member of set A is also a member of set B. Figure 1.1 presents estimated accuracy as a function of L for

sites in set A imputed with Beagle. As might be expected, we observed greatest variation among accuracy estimates for small L (Figure 1.1). We removed sites with L<30 from our analysis, leaving us with 9737 sites (set C) to analyze. We selected a filter threshold of 30 somewhat arbitrarily but opted for a moderate-stringency filter to avoid removing a large subset of sites from our analysis.

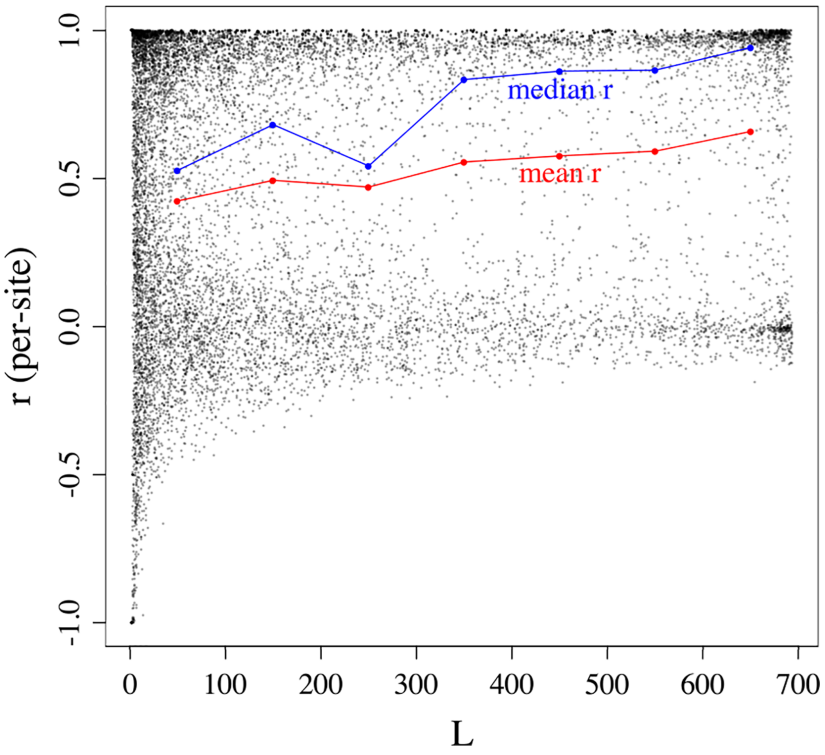


Figure 1.1 Estimates of accuracy as a function of *L* for 13028 sites imputed with Beagle. Imputation accuracies were estimated using the sample Pearson correlation coefficient, *r*. The sample Pearson correlation is a function of two vectors, both of length *L*. Figure 1.1. presents estimated accuracy as a function of *L* for set A sites (n=13028). The range of *L* is divided into a series of seven equally sized bins (i.e. $0 < L \leq 100$, $100 < L \leq 200$, ..., $600 < L \leq 700$). Accuracy estimates were divided into bins according to their corresponding values of *L*. Bin means and medians are presented as red and blue points, respectively.

Figure 1.2 summarizes and compares the accuracy of Beagle and glmnet imputation at the site and individual level. Both Beagle and glmnet produced

bimodal distributions of per-site accuracies, with median per-site imputation accuracies of 0.76 and 0.82, respectively (Figure 1.2B). We argue that this bimodality results from an overrepresentation of low-frequency variants, a hallmark of HTS-derived datasets. Both methods produced left-skewed distributions of per-individual Pearson correlations, with nearly identical medians (0.991 and 0.992 for Beagle and glmnet, respectively; Figure 1.2D).

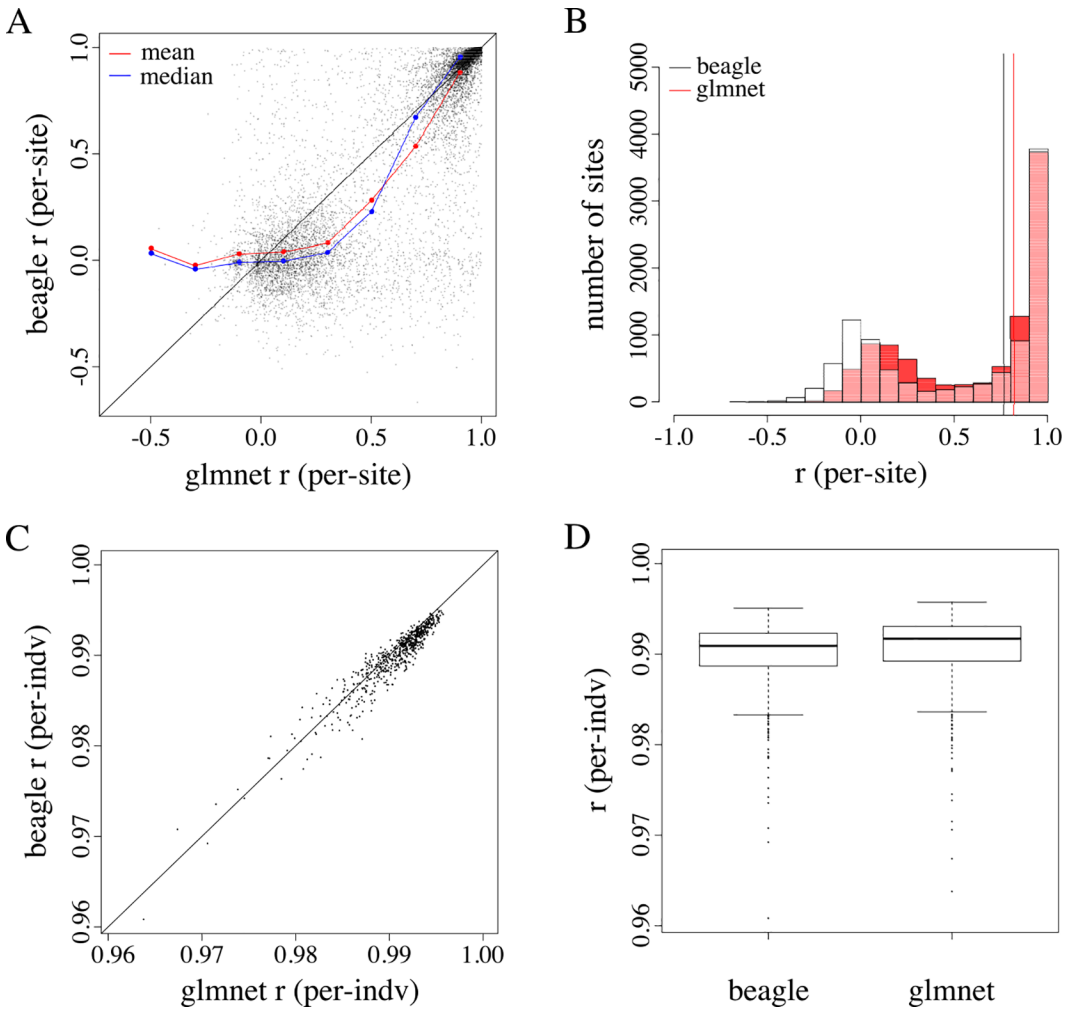


Figure 1.2. A summary and comparison of per-site and per-individual imputation accuracy from Beagle and glmnet imputation.
(A and B) The *x*- and *y*-axes report estimates of imputation accuracy for glmnet and Beagle, respectively. Each point represents the estimated accuracy for a single site

(A) and individual (B). (C) Both Beagle and glmnet produced bimodal distributions of per-site accuracies, with median per-site imputation accuracies of 0.76 (black vertical line) and 0.82 (red vertical line), respectively. (D) Both methods produced left-skewed distributions of per-individual accuracies, with median per-individual accuracies of 0.991 and 0.992 for Beagle and glmnet, respectively.

Proportion of missing data and read depth

We examined the effect of the proportion of missing data on imputation accuracy at the site and individual level (Figure 1.3). As might be expected, we observed a decline in imputation accuracy as the level of missing data increased. Beagle appears to show greater sensitivity to levels of missing data relative to glmnet, particularly when the proportion of missing data at a site falls within the (0.1, 0.5] interval (Figure 1.3A and 3B). We observed essentially no difference between the two imputation methods when examining accuracy at the individual level (Figure 1.3B).

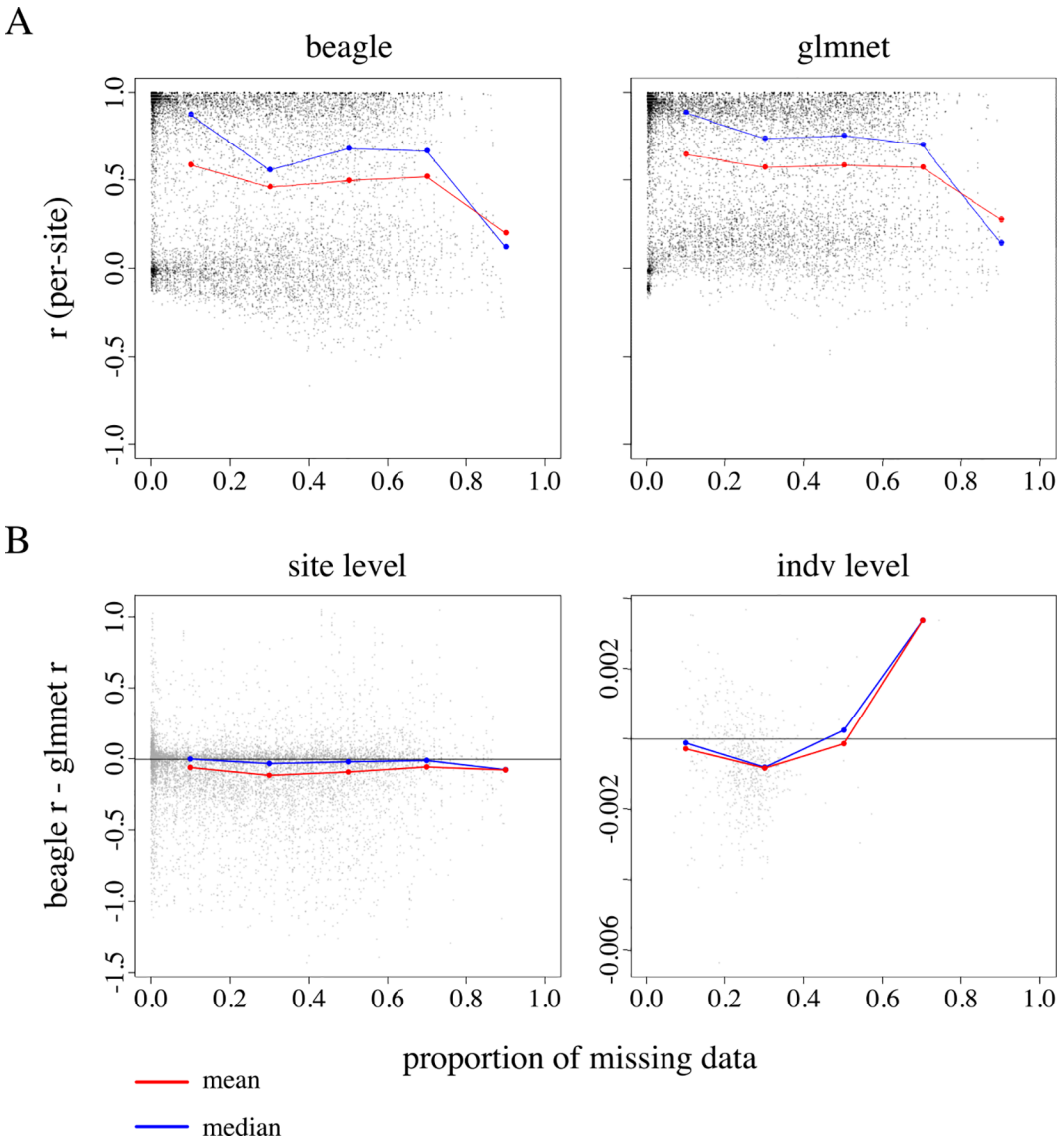


Figure 1.3. Per-site and per-individual imputation accuracy as a function of missing data and median read depth.
(A) Beagle and glmnet imputation accuracy as a function of missing data for sites in set C (n = 9737). (B) The x- and y-axis display the proportion of missing data and the accuracy difference between Beagle and glmnet at the site and individual level. The range of x is divided into ten-equally sized bins (i.e. $0.00 < x \leq 0.10$, $0.10 < x \leq 0.20$, ..., $0.90 < x \leq 1.00$), and accuracy differences are divided into bins according to levels of missing data. Bin means and medians, summarizing the data within each bin, are displayed as red and blue points, respectively. Points falling on the black vertical line at y = 0 indicate no observed accuracy difference between Beagle and glmnet imputation. Points falling below y = 0 represent cases where glmnet imputes with higher accuracy relative to Beagle.

Minor allele frequency

We estimated the minor allele frequency (MAF; the minor allele at a site could be either the reference or alternative allele listed in the VCF file) at all 20205 sites using the sample of 694 individuals from the GG population. Figure 1.4 presents per-site *r* as a function of estimated MAF for the 9737 sites in set C. We divided the range of *x* into five-equally sized bins (i.e. $0.00 < x \leq 0.10$, $0.10 < x \leq 0.20$, ..., $0.40 < x \leq 0.50$), and summarized accuracy values within each frequency bin using the mean and median (Figure 1.4). We observed a decrease in accuracy as MAF decreased and greatest variance in low-frequency bins (Figure 1.4 left and middle panel). These two trends are consistent with previous results suggesting that sites harboring rare alleles are more difficult to impute accurately relative to sites harboring more common alleles [3]. Glmnet appears to impute with slightly higher accuracy than Beagle at all MAF bins (Figure 1.4 right panel).

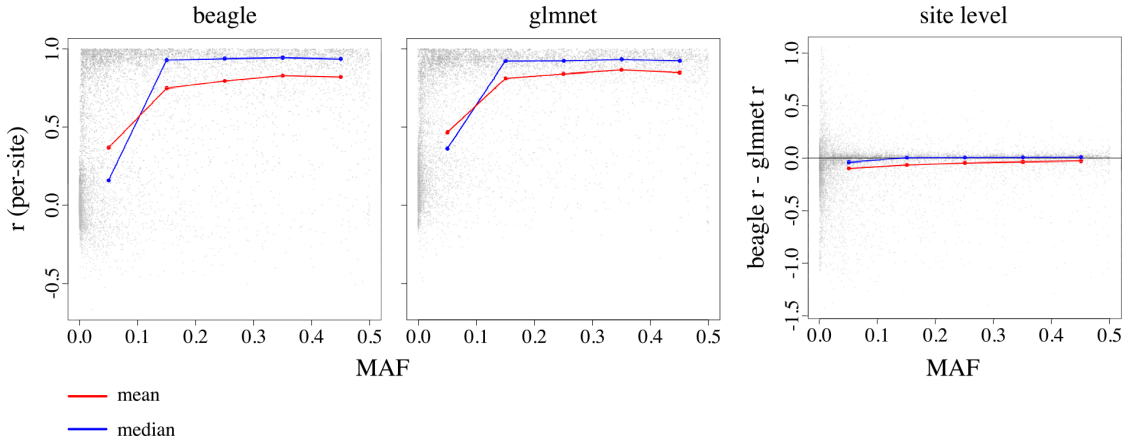


Figure 1.4. Imputation accuracy as a function of MAF
The left and middle panels show per-site accuracy of Beagle and glmnet as a function of (estimated) MAF. The right-most panel shows the difference in accuracy between Beagle and glmnet at each site as a function of MAF. We observed the

greatest difference in accuracy at low-frequency variants. Low-frequency variants were imputed with high variance.

Reference Panel Size and Composition

We next investigated the effect of reference panel composition on imputation accuracy (Figure 1.5). Figure 1.5 summarizes Beagle and glmnet imputation accuracy in a sample of 2490 individuals from the C1 population for genotypes imputed with a reference panel of 694 and 80 individuals (Figure 1.5A). [19] reported considerable increases in Beagle's imputation accuracy with increased reference panel size across all minor allele frequencies, with the greatest increase at low-frequency variants. We, however, observed essentially no difference in the median per-marker *r* when imputing with the larger reference panel (Figure 1.5A). Sites with a MAF ≤ 0.01 appeared to benefit the most when imputing with a larger reference panel, but gains in accuracy were small (Figure 1.5B). We observed modest gains in mean accuracy across all levels of missing data when imputing with the larger reference panel (Figure 1.5C). Overall, Beagle and glmnet imputed missing genotype with similar accuracies regardless of the reference panel used. Beagle required a slightly longer runtime relative to glmnet (Table 1.1).

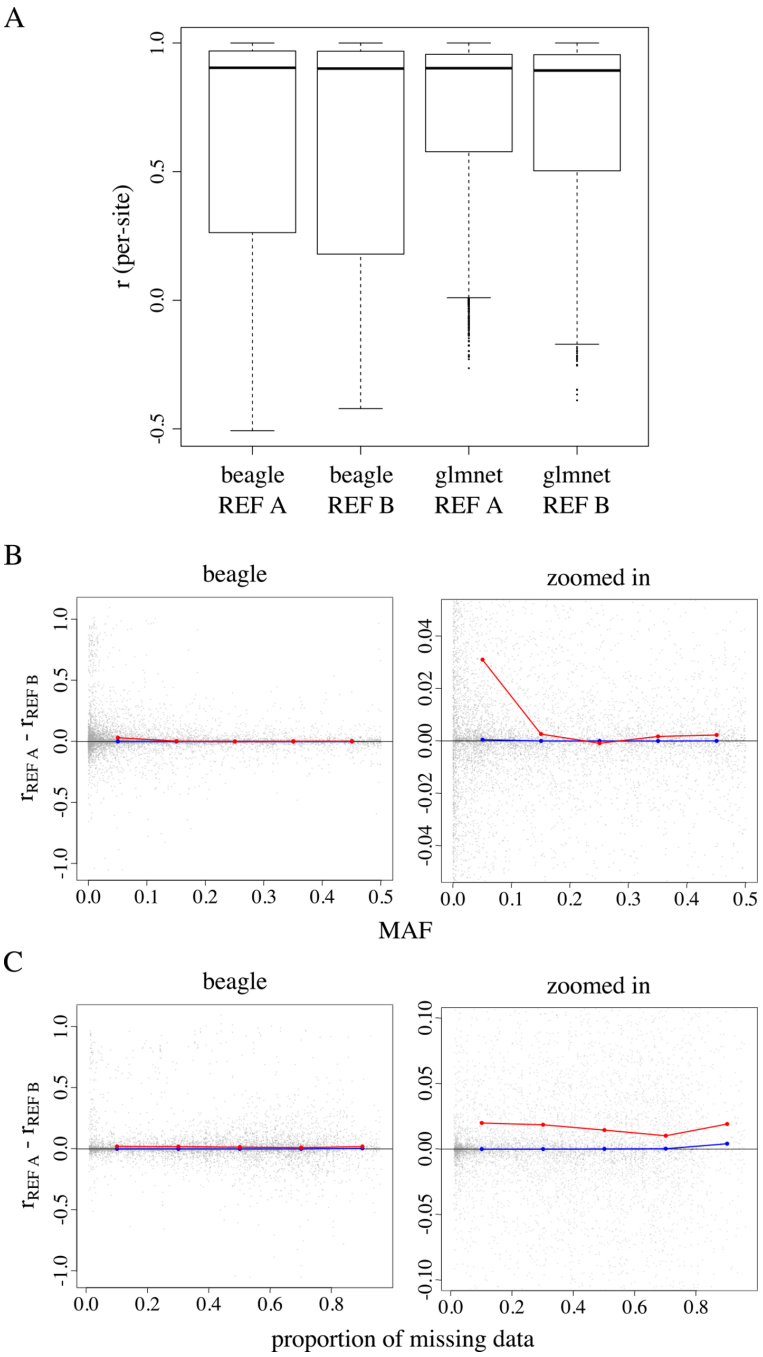


Figure 1.5. The accuracy difference between reference panel A and panel B as a function of MAF and proportion of missing data for 11535 sites.
(A) Genotypes in a sample of 2490 C1 individuals were imputed using two different reference panels: reference panel A, comprised of 694 phased GG individuals, and reference panel B, comprised of 80 phased individuals listed as progenitors of the C1 population. (B and C) Points falling on the black vertical line at $y = 0$ indicate no observed accuracy difference when imputing with reference panel A or B. Points falling below $y = 0$ represent cases where Beagle imputes with higher accuracy when using reference panel B relative to imputing with reference panel A.

Table 1.1. A summary of Beagle and glmnet's computation cost (in seconds) and median per-site and per-individual accuracy under scenario 1, 2, and 3. (Top) We calculated the mean computation time across the 10 folds of each cross-validation experiment. (Middle) We calculated the median *r* across sites and reported this as a scalar summary of imputation accuracy in that cross-validation experiment. (Bottom) We calculated the median *r* across individuals and reported this as a scalar summary of imputation accuracy in that cross-validation experiment

Mean computation time	Scenario 1 (seconds)	Scenario 2 (seconds)	Scenario 3 (seconds)
Beagle	2249.6	63713.5	43935.4
Glmnet	12477.86	56295.45	34551.17

Median per-site r	Scenario 1 (percent)	Scenario 2 (percent)	Scenario 3 (percent)
Beagle	76.48	90.37	90.05
Glmnet	81.94	90.21	89.31

Median per-individual r	Scenario 1 (percent)	Scenario 2 (percent)	Scenario 3 (percent)
Beagle	99.17	99.34	99.36
Glmnet	99.09	99.30	99.29

DISCUSSION

Imputation accuracy was calculated as the correlation between the observed genotype dosage (estimated from allelic count data in the AD subfield of the VCF file) and the imputed genotype dosage. We note that to obtain true measures of imputation accuracy, the imputed genotype dosage should be correlated with the true genotype, rather than the observed genotype dosage. Unfortunately, true genotypes are not known and observed genotype dosages must be used instead. The accuracy based on correlation to the observed genotype dosages under-estimates the true imputation accuracy in two ways. First, there is error associated with the observed genotype dosage (resulting from sequencing errors, alignment errors, etc.)

that reduces the correlation. Second, the observed genotype dosage of individual *i*, at site *j* were computed using one source of information -- the observed sequence data from individual *i*, at site j. The imputed genotype dosage from Beagle and glmnet, in contrast, were computed using a multi-sample LD approach. Multi-sample LD methods infer the genotype dosage of individual *i*, at site *j* by jointly analyzing data from multiple individuals in the sample, at site *j* and at nearby sites (i.e. information regarding LD). The use of information from multiple individuals and patterns of LD has been shown to lead to significant improvements in genotype-calling accuracy for low-depth sequence data (for an example, see [20]).

Using a set of validation genotypes at biallelic SNPs on chromosome 5, we found that Beagle and glmnet impute missing variants with similar accuracies. When comparing the two methods at the site level, glmnet appears to impute with (moderately) higher accuracy relative to Beagle, regardless of levels of missing data. We, however, observe little difference between the two methods when measuring accuracy at the individual level (Figure 1.3B). We observed the greatest difference in accuracy between the two methods in scenario 1 (imputation guided by no reference panel). Differences, however, were only moderate, suggesting that 1) human-tailored imputation algorithms can produce relatively accurate genotype estimates when applied to datasets derived from non-human organisms and 2) these algorithms, which were developed and optimized to impute ascertained variants, appear amenable for imputation of variants discovered via an HTS methods such as GBS.

The unique aspects of the datasets derived from a non-human organism, such as cassava, and HTS methods, such as GBS, do not affect Beagle's imputation accuracy in ways we do not understand or expect. For instance, we observed a decrease in imputation accuracy as MAF decreased (Figure 1.4 left), consistent with previous results suggesting that sites harboring rare alleles are more difficult to impute accurately relative to sites harboring more common alleles [3]. Results also indicate that the Beagle algorithm is robust to deviations from the HWE assumption that underlies the Beagle algorithm. HWE is violated in domesticated species, which have undergone generations of controlled mating and directional selection.

The modest difference in imputation accuracy between Beagle and glmnet was in some ways unexpected, largely because the two algorithms employ contrastingly different approaches to modeling genotype data. Glmnet does not attempt to directly relate observed correlation patterns to any underlying biological process, whereas Beagle specifies a statistical model for the biological aspect of the problem – namely, the haplotypes that generated the observed LD structure. Both algorithms leverage data at a subset of markers to impute missing genotypes at a particular locus, but they employ very different subset selection strategies. Glmnet selects markers solely on measures of pairwise correlation, ignoring locus order and spacing. Beagle, in contrast, focuses on a small number of nearby markers when imputing missing genotypes at a particular site (localized haplotype-cluster model). Correlation between markers is a localized phenomenon; that is, there tends to be less LD between loci that are far apart than between loci that are close together. The apparent correlations observed between distant markers are largely statistical

artifacts, i.e. noise introduced by sampling variation. While glmnet and Beagle produced similar results in our cross-validation experiments, we reason that there are situations in which Beagle will outperform glmnet (e.g. when levels of spurious associations between distant markers is high relative to true levels of LD). In addition to decreased sensitivity to spurious associations between distant markers, probabilistic, phasing methods, such as Beagle, offer additional benefits, such as providing phased haplotypes and measures of imputation accuracy estimated from posterior genotype probabilities.

In scenario 2 and 3, we used a sample of 2490 C1 individuals to compare the accuracy of genotype imputation with a cosmopolitan reference panel (reference panel A) and a best-match panel (reference panel B). Reference panel A consists of 694 individuals, a subset of who are list as C1 parents in pedigree records (n = 78). Reference panel B, in contrast, consists entirely of individuals listed as C1 parents in pedigree records (n = 80). The set of C1 parents in panel A is a subset of panel B. We found that imputation using reference panel A and B resulted largely in similar imputation accuracies across sites. We find this reassuring for two reasons: 1) the 617*2 haplotypes from the non-parental individuals in reference panel A appear to serve as good proxies for the haplotypes of the two C1 parents that are present in panel B but absent in panel A and 2) adding 'extraneous' haplotypes to the reference panel appears to introduce little error to the imputation procedure, consistent with previous observations made by those in the human genetics community [3]. Imputation with reference panel A required more computation time relative to imputation with panel B (by approximately 1.5X). In practice, however, the task of

constructing a best-match reference panel is considerably more challenging and computationally expensive than the one presented here. We reason that a cosmopolitan reference panel is a good fallback choice when the optimal panel composition is unclear and if one has the computational resources to employ a large reference panel for imputation.

REFERENCES

[1] Encode Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2013.

[2] Y. Li, C. Sidore, H. M. Kang, M. Boehnke, and G. R. Abecasis, "Low-coverage sequencing: Implications for design of complex trait association studies," *Genome Res.*, vol. 21, no. 6, pp. 940–951, 2011.

[3] B. Howie, J. Marchini, M. Stephens, and a. Chakravarti, "Genotype Imputation with Thousands of Genomes," *G3: Genes|Genomes|Genetics*, vol. 1, no. 6, pp. 457–470, 2011.

[4] B. L. Browning and Z. Yu, "Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies," *Am. J. Hum. Genet.*, vol. 85, no. 6, pp. 847–861, 2009.

[5] J. E. Rutkoski, J. Poland, J.-L. Jannink, and M. E. Sorrells, "Imputation of unordered markers and the impact on genomic selection accuracy.," *G3 (Bethesda).*, vol. 3, no. 3, pp. 427–39, 2013.

[6] R. J. Elshire, J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and

S. E. Mitchell, “A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species,” *PLoS One*, vol. 6, no. 5, 2011.

[7] J. C. Glaubitz, T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire, Q. Sun, and E. S. Buckler, “TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline,” *PLoS One*, vol. 9, no. 2, 2014.

[8] M. D. Wolfe, I. Y. Rabbi, C. Egesi, M. Hamblin, R. Kawuki, P. Kulakow, R. Lozano, D. P. del Carpio, P. Ramu, and J.-L. Jannink, “Genome-wide association and prediction reveals the genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement,” *Plant Genome*, pp. 1–248, 2016.

[9] H. Ceballos, C. A. Iglesias, J. C. Pérez, and A. G. O. Dixon, “Cassava breeding: Opportunities and challenges,” *Plant Mol. Biol.*, vol. 56, no. 4, pp. 503–516, 2004.

[10] J. B. Endelman and J.-L. Jannink, “Shrinkage estimation of the realized relationship matrix.,” *G3 (Bethesda).*, vol. 2, no. 11, pp. 1405–13, 2012.

[11] J. Poland, J. Endelman, J. Dawson, J. Rutkoski, S. Y. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sanchez-Villeda, M. Sorrells, and J. L. Jannink, “Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing,” *Plant Genome*, vol. 5, no. 3, pp. 103–113, 2012.

[12] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization Paths for Generalized Linear Models via Coordinate Descent.,” *J. Stat. Softw.*, vol. 33, no. 1, pp. 1–22, 2010.

[13] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, *The elements of statistical*

learning: data mining, inference and prediction, vol. 27, no. 2. 2005.

[14] S. R. Browning and B. L. Browning, "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.," *Am. J. Hum. Genet.*, vol. 81, no. 5, pp. 1084–97, 2007.

[15] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2. pp. 257–286, 1989.

[16] S. R. Browning, "Multilocus association mapping using variable-length Markov chains.," *Am. J. Hum. Genet.*, vol. 78, no. 6, pp. 903–13, 2006.

[17] M. P. L. Calus, a C. Bouwman, J. M. Hickey, R. F. Veerkamp, and H. a Mulder, "Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications.," *Animal*, pp. 1–11, 2014.

[18] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011.

[19] B. L. Browning and S. R. Browning, "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals," *Am. J. Hum. Genet.*, vol. 84, no. 2, pp. 210–223, 2008.

[20] 1000 Genomes Project Consortium, others, and W. Africa, "A map of human genome variation from population-scale sequencing.," *Nature*, vol. 467, no.

7319, pp. 1061–73, 2010.

CHAPTER 2

A STATISTICAL FRAMEWORK FOR DETECTING MISLABELED AND CONTAMINATED SAMPLES USING SHALLOW-DEPTH SEQUENCE DATA[2]

ABSTRACT

Researchers typically sequence a given individual multiple times, either re-sequencing the same DNA sample (technical replication) or sequencing different DNA samples collected on the same individual (biological replication) or both. Before merging the data from these replicate sequence runs, it is important to verify that no errors, such as DNA contamination or mix-ups, occurred during the data collection pipeline. Methods to detect such errors exist but are often ad hoc, cannot handle missing data and several require phased data. Because they require some combination of genotype calling, imputation, and haplotype phasing, these methods are unsuitable for error detection in low- to moderate-depth sequence data where such tasks are difficult to perform accurately. Additionally, because most existing methods employ a pairwise-comparison approach for error detection rather than joint analysis of the putative replicates, results may be difficult to interpret. We introduce a new method for error detection suitable for shallow-, moderate-, and high-depth sequence data. Using Bayes Theorem, we calculate the posterior probability distribution over the set of relations describing the putative replicates and infer which of the samples originated from an identical genotypic source. Our

[2] A. W. Chan, A. L. Williams, and J.-L. Jannink, "A statistical framework for detecting mislabeled and contaminated samples using shallow-depth sequence data," *BMC Bioinformatics*, pp. 1–14, 2018.

method addresses key limitations of existing approaches and produced highly accurate results in simulation experiments. Our method is implemented as an R package called BIGRED (Bayes Inferred Genotype Replicate Error Detector), which is freely available for download: https://github.com/ac2278/BIGRED.

INTRODUCTION

A researcher may choose, for a number of reasons, to sequence an individual multiple times, performing technical replication, biological replication, or both. Because sequencing experiments involve many steps and errors can occur during any part of the workflow, one motivation for sequencing an individual more than once is to allow researchers to compare these replicates, identify outlier samples, and evaluate how well a sequencing pipeline is executed. This is particularly important for plant breeders, as they require ongoing estimates of their program's error rates. Further discussion of reasons for intentional replication appear elsewhere [1]. In short, the three aspects of replication—sequencing read depth, technical replication, and biological replication—each play different roles in mitigating errors that are introduced in the experimental pipeline. Increasing sequencing read depth allows for improved variant calling while technical and biological replicates allow for optimization of bioinformatic filters [1]. Replication can also arise unintentionally as a result of human error or naming inconsistencies, and it is in a researchers best interest to make full use of the data, merging the replicate records rather than discarding them.

Before merging the data from biological or technical replicates or using them to inform quality filter thresholds, it is important to verify that no erroneous samples exist among the putative replicates (i.e. verify that all putative replicates derived from an identical individual). Existing methods for error detection include performing pairwise identity-by-state and –by-descent estimation [2], calculating the correlation between pairs of samples, and examining a heat map of a realized genomic relationship matrix. These approaches require some combination of genotype calling, imputation, and haplotype phasing, making them unsuitable for low- to moderate-depth, high-throughput sequence (HTS) data [3]. And because these methods employ a pairwise-comparison approach for error detection rather than joint analysis of the samples, results may be inconsistent when more than two replicates exist. To illustrate, the general protocol for heat map analysis involves starting off with some collection of sequenced samples (including the replicates of interest), calling genotypes, filtering based on percent missing, imputing missing genotypes, calculating the additive genomic relationship matrix, and finally plotting a heat map of the putative replicates. This method can work well on deeply sequenced samples, but complications arise when applying this method to shallow-depth sequence data. Firstly, it requires genotype calling, which is difficult to do accurately when we have low read depth. Secondly, it requires imputation, raising issues in regards to reference panel and imputation method selection. Furthermore, results from imputation vary depending on which samples were jointly imputed, which in turn, affects downstream analyses that use the imputed data. Finally, a third limitation of this method—common among existing error detection

methods—is that it relies on pairwise comparisons of the putative replicates, rather than joint analysis of the replicates. For example, suppose we have three putative replicates, A, B, and C. It is possible that A and B are highly correlated, A and C are highly correlated, but B and C are only moderately correlated. In situations such as this, deciding if all three samples are replicates is not straightforward.

Considering these issues, we propose a method that addresses key limitations of existing approaches. The proposed method detects errors by estimating the conditional posterior probability of all possible relationships among the putative replicates (Figure 2.1). We call our algorithm BIGRED (Bayes Inferred Genotype Replicate Error Detector). BIGRED requires no genotype calling, imputation, or haplotype phasing, making it a suitable tool for studies relying on shallow-depth HTS data. We examined the effect of read depth, the number of sites analyzed (L), and minor allele frequency (MAF) at the L sites on algorithmic performance, using both real and simulated data. In this paper, we used BIGRED as a tool to verify reported replicates; however, we also envision individuals using our algorithm to test unreported but suspected replicates. Under this scheme, researchers would use some initial screening method, such as examination of the genomic relationship matrix, to identify cryptic replicates among their collection of samples and then test these suspected replicates using BIGRED.

MATERIALS AND METHODS

The Proposed Method

We describe the proposed method using a case study, individual I011206 from the Next Generation (NEXTGEN) Cassava Breeding Project [4]. I011206 is recorded to have been sequenced $k = 3$ times by NEXTGEN (Appendix Figure 2.1). We index the putative replicates using the variable d. The task is to verify that samples $d = 1$, $d = 2$, and $d = 3$ are in fact replicates of the same individual, checking all possible combinations of replicate and non-replicate status. We know that the DNA samples from these three runs can be related in one of five possible ways (Figure 2.1):

1. All three samples originate from one source;
2. Samples $d = 1$ and $d = 2$ originate from one source while $d = 3$ originates from a different source;
3. Samples $d = 1$ and $d = 3$ originate from one source while $d = 2$ originates from a different source;
4. Samples $d = 2$ and $d = 3$ originate from one source while $d = 1$ originates from a different source;
5. All three samples originate from different sources.

We use "source vectors" to represent these relations and enumerate all possible source vectors for $k = 3$ on the right panel of Figure 2.1. By convention: (1) source vectors are *labeled* vectors, e.g., the first, second, and third element of a given source vector describes the status of sample $d = 1$, $d = 2$, and $d = 3$, respectively, and (2) the first element of a source vector always takes on the value 1. Vector elements with the same value are indicated to be from the same source.

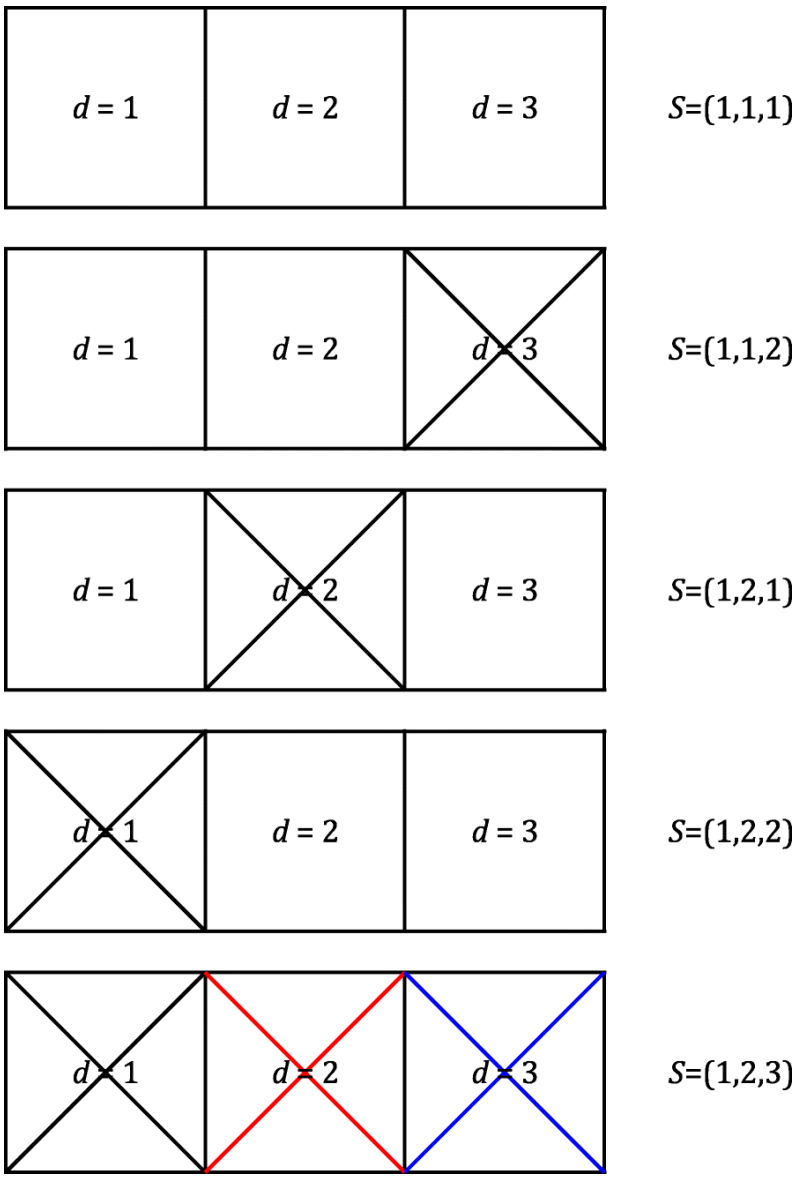


Figure 2.1. The set of relations describing the three putative replicates of an individual and the corresponding source vectors.
BIGRED calculates the posterior probability distribution over the set of relations describing the putative replicates and infers which of the samples originated from an identical genotypic source. The source vector S = (1,2,1) represents the scenario where sample d = 1 and d = 3 originate from an identical source. Crossed out boxes represent samples without any replicate.

BIGRED detects errors by estimating the conditional posterior probability of each

source vector S, given:

1. Estimates of population allele frequency at *L* randomly sampled biallelic sites, sampled at the genome-wide level and
2. The *k* putative replicates' allelic depth (AD) data at the *L* sites. A site is only sampled if each putative replicate has at least one read at that site.

We make three simplifying assumptions:

1. The species is diploid;
2. Each polymorphic site harbors exactly two alleles, allele *A* and allele *B*, i.e. all polymorphisms are biallelic;
3. Sites are independent. BIGRED allows the user to specify a minimum distance, in base pairs, between any two sampled sites. The user may also filter sites based on linkage disequilibrium, although this is not a functionality of BIGRED.

Defining a likelihood function for G

Let $X_d^{(v)}$ and $G_d^{(v)}$ denote the observed AD data and the underlying (unknown) genotype at site *v* for putative replicate *d*, respectively. The AD data records the observed counts of allele *A* and *B* at site *v* for sample *d*: $X_d^{(v)} = (n_A^{(v,d)}, n_B^{(v,d)})$. Given observed data $X_d^{(v)}$ and fixed sequencing error rate *e*, we compute the likelihood for genotype $G_d^{(v)} = g$ at site *v* for sample *d* using a binomial model as follows, where $g \in AA, AB, BB$:

$$P\left(X_d^{(v)} \middle| G_d^{(v)} = g, e\right) = \binom{n_A^{(v,d)} + n_B^{(v,d)}}{n_B^{(v,d)}} (1 - p_B)^{n_A^{(v,d)}} (p_B)^{n_B^{(v,d)}}$$

$$p_B = \begin{cases} e, & \text{when } g = 0 \text{ or } AA \\ 0.50, & \text{when } g = 1 \text{ or } AB \\ 1 - e, & \text{when } g = 2 \text{ or } BB \end{cases} \quad (1)$$

Defining a likelihood function for S

We walk through the procedure of defining the likelihood function for *S* when *k* = 3, continuing with individual I011206 as an example:

1. Enumerate all possible source vectors of length *k* = 3 (Figure 2.1).
2. Enumerate all *labeled* genotype vectors consistent with each source vector (Figure 2.2). For instance, there are three genotype vectors consistent with source vector *S* = (1,1,1): (*AA*, *AA*, *AA*), (*AB*, *AB*, *AB*), and (*BB*, *BB*, *BB*). There are nine genotype vectors consistent with *S* = (1,1,2): (*AA*, *AA*, *AA*), (*AA*, *AA*, *AB*), (*AA*, *AA*, *BB*), (*AB*, *AB*, *AB*), (*AB*, *AB*, *AA*), (*AB*, *AB*, *BB*), (*BB*, *BB*, *BB*), (*BB*, *BB*, *AA*), and (*BB*, *BB*, *AB*).
3. Define a likelihood function for *S* as a function of genotype likelihoods, defined previously in Equation 1:

$$
\begin{aligned}
P\left(X^{(v)}\middle|S\right) &= \sum_{G^{(v)}} P\left(X^{(v)}, G^{(v)}\middle|S\right) \\
&= \sum_{G^{(v)}} P\left(X^{(v)}\middle|G^{(v)}\right)P\left(G^{(v)}\middle|S\right) \\
&= \sum_{G^{(v)}} \left[\prod_{d=1}^{k} P\left(X_d^{(v)}\middle|G_d^{(v)}\right)\right] P\left(G^{(v)}\middle|S\right)
\end{aligned}
\tag{2}
$$

The function $P\left(G^{(v)}\middle|S\right)$ is the probability that the k samples have genotype vector $G^{(v)} = (G_{d=1}^{(v)}, G_{d=2}^{(v)}, \ldots, G_{d=k}^{(v)})$ given that source vector S describes how the k samples are related. We define $P\left(G^{(v)}\middle|S\right)$ using the (user-supplied) population allele frequency of allele B at site v and assuming Hardy-Weinberg Equilibrium (HWE; Figure 2.2). For samples that are encoded as identical in source vector S, we treat their genotypes as a single observation and all non-identical genotypes are modeled as independent (Figure 2.2).

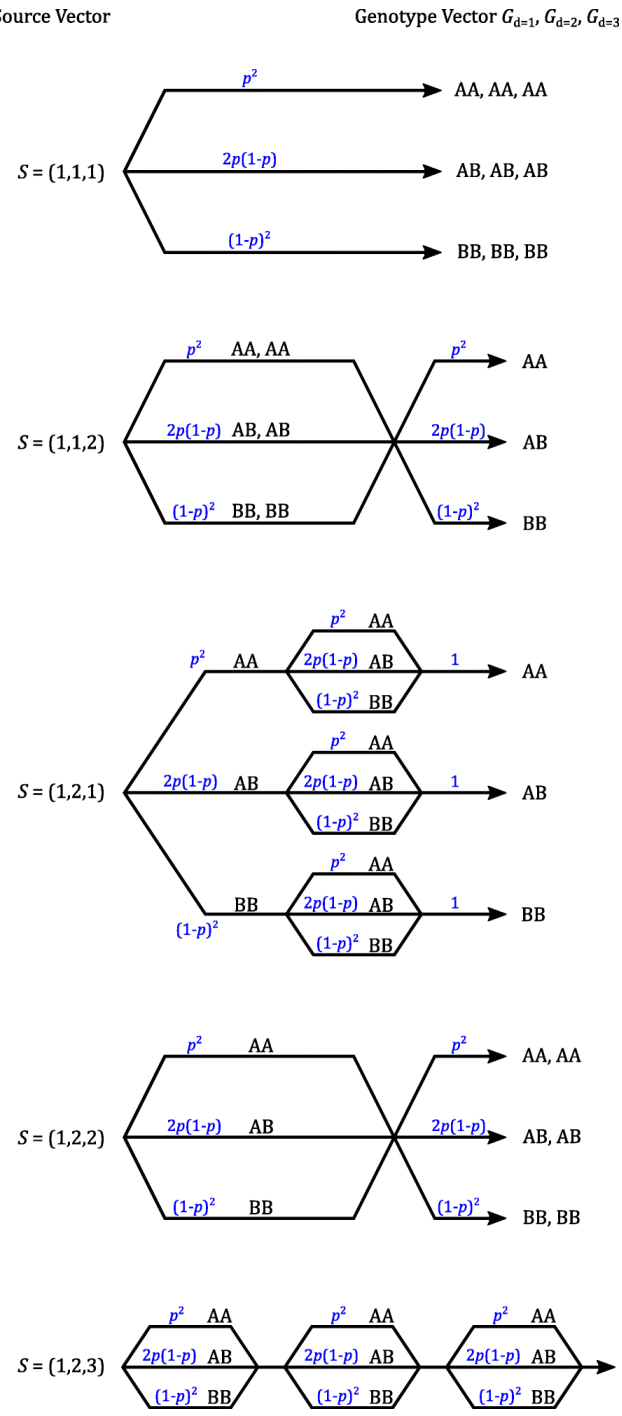


Figure 2.2. Defining $P\left(G^{(v)}\middle|S\right)$ for k = 3.
We first enumerate all possible source vectors of length k = 3 (left) then enumerate all *labeled* genotype vectors consistent with each source vector (right). Each path in a given tree corresponds to a genotype vector given source vector S. For instance, if the three samples are related by source vector (1,1,2), the genotype vector can take one of nine values. We compute the probability of each genotype vector (given S) by traversing each path and taking the product of the probabilities associated with the

edges of the path. Note that genotype vectors not consistent with S have probability zero (we omit these paths from the figure). Edge probabilities are defined using user-supplied, population allele frequencies and assuming HWE.

Estimating $P(S|X)$

Once we compute $P\left(X^{(v)}\middle|S\right)$ at all L sites, we compute $P(S|X)$ jointly across all L sites using Equation 3 and assuming a uniform prior on S:

$$\prod_{v=1}^{L} P\left(X^{(v)}\middle|S\right) = P(X|S)$$

$$P(S|X) = \frac{P(X|S)P(S)}{\sum_S [P(X|S)P(S)]} \tag{3}$$

One may wish to compare the posterior probability of two assignments of S, and when doing so via the posterior odds-ratio, both the denominator and $P(S)$ cancel from the two posteriors (since the denominator acts as a normalizing constant and we assume a uniform prior on S). The ratios of the posteriors are, therefore, equal to the ratios of the likelihoods.

Evaluating BIGRED

We examined how changes in mean read depth, L, and MAF at the L sites affect the accuracy of BIGRED. For simulation experiments, we used a fixed sequencing error rate of 0.01 and sampled sites such that no two sites fell within 20 kb from one another. In addition to accuracy, we evaluated the sensitivity of the algorithm. We used high-depth whole-genome sequence (WGS) data from 241

Manihot esculenta individuals to simulate a series of data sets. Filtering the data (e.g., removing sites with extremely low minor allele frequency and discarding regions prone to erroneous mapping) should be done prior to applying BIGRED to remove potentially spurious variants. We refer the reader to the section "Alignment of reads and variant calling of cassava haplotype map (HapMapII)" of [5] for a description of how the data was generated and the quality filters applied.

The data

The WGS data consist of both AD data and called genotypes for 241 individuals. To detect the presence of any population structure, we performed principal component analysis (PCA) using the called genotypes for the 241 individuals. We generated a pruned subset of SNPs that are in approximate linkage equilibrium with each other and then performed a PCA using this pruned subset of SNPs (Figure 2.3). We performed LD-based SNP pruning and PCA using R packages SNPRelate() and gdsfmt() with a LD threshold of 0.40 [6]. The 241 individuals clustered into roughly three groups. The 206 individuals shown in orange represent cultivated cassava. We used these 206 individuals to estimate population allele frequencies at sites and 15 individuals, previously found to be genetically distinct [7], to simulate AD data for experiments. We limited our simulation experiments to these 15 members to ensure that all individuals truly represent distinct genotypes rather than only nominally distinct.

Simulation experiments to evaluate the impact of mean read depth and MAF on accuracy

We first evaluated the effect of mean read depth λ and MAF on the algorithm's accuracy, holding L constant at 1000 sites. We outline the procedure to simulate AD data for the scenario where $k = 3$ and $S = (1,2,1)$:

1. Enumerate all possible pairs of genotypes, where order does not matter ($n = 15(14) = 210$).
2. Sample one genotype pair.
3. Randomly assign the status 'source 1' to one of the two genotypes. Assign the remaining genotype 'source 2' status.
4. Randomly sample $L = 1000$ sites (genome-level) with a specified MAF.
5. Simulating $X_{d=1}^{(v)}$: Sample Y alleles (with replacement) from the pool of allele reads belonging to source 1 at that site, where $Y \sim Poisson(\lambda)$.
6. Simulating $X_{d=2}^{(v)}$: Sample Y alleles (with replacement) from the pool of allele reads belonging to source 2 at that site, where $Y \sim Poisson(\lambda)$.
7. Simulating $X_{d=3}^{(v)}$: Sample Y alleles (with replacement) from the pool of allele reads belonging to source 1 at that site, where $Y \sim Poisson(\lambda)$.
8. Feed the algorithm the simulated AD data and the population allele frequency of allele B at the L sites.
9. Record the conditional posterior probability of $S = (1,2,1)$.
10. Repeat steps 2 through 9, 100 times. When repeating step 2, only sample from those genotype pairs that have not been sampled previously.

Note that evaluating scenario S = (1,2,1) is equivalent to evaluating scenarios S = (1,1,2) and S = (1,2,2). We performed a full factorial experiment for the source vectors associated with k = 2, k = 3, and k = 4, where λ = {1,2,3,6,15} and where we sampled sites with a given MAF falling in one of five possible intervals (0.0,0.1], (0.1,0.2], (0.2,0.3], (0.3,0.4], and (0.4,0.5]. Note that in these simulation experiments, all putative replicates of a given individual had identical mean read depths. We later tested the scenario where mean read depths varied among the samples.

Simulation experiments to evaluate the impact of L on accuracy

To assess the impact of L on accuracy, we repeated simulation experiments for S = (1,2,1) and S = (1,2,3), sampling sites with MAFs falling in (0.2,0.3] and testing seven values of L: 50, 100, 250, 500, 1000, 2000, and 5000.

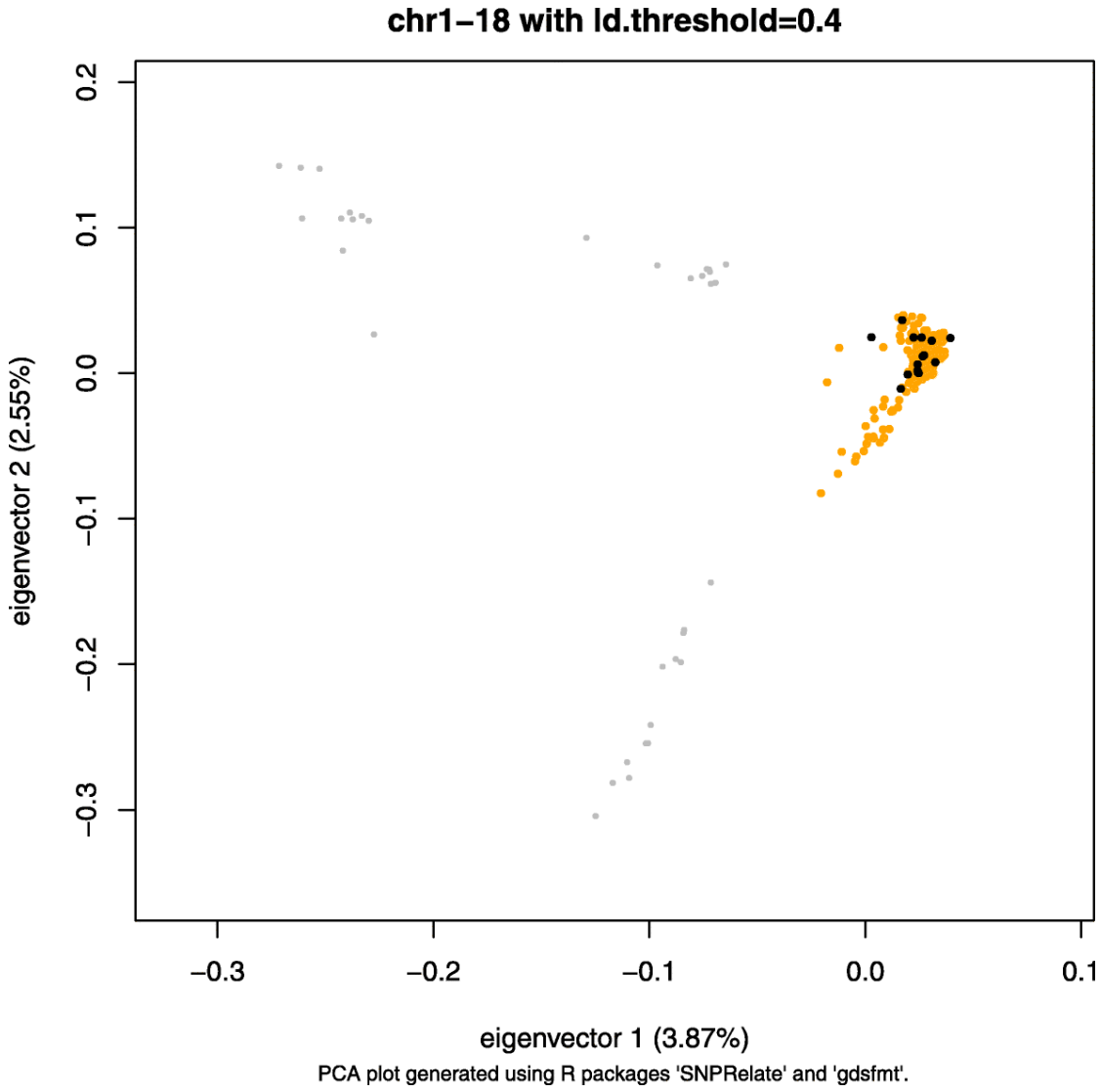


Figure 2.3. PCA on 241 *Manihot esculenta* genotypes, using a subset of SNPs in approximate linkage equilibrium.
The x-axis and y-axis in this figure represents the first and second eigenvector, respectively. The 241 individuals clustered into roughly three groups. We used cultivated cassava (orange and black) to evaluate BIGRED in simulation experiments. We used 15 individuals (black) to simulate AD data and all 206 (orange and black) individuals to estimate population allele frequencies at sites.

Simulation experiments to evaluate BIGRED's sensitivity

We next evaluated the algorithm's sensitivity by simulating the scenario

where S = (1,1) and corrupting (i.e., contaminating) p percent of sites in sample d = 2

with a second, randomly sampled genotype source. We tested five values of p (10%, 20%, 30%, 40%, 50%) at five mean depths (1x, 2x, 3x, 6x, and 15x). We repeated this procedure 100 times for each depth and p combination.

Simulation experiments to evaluate the scenario where mean read depths vary among the k putative replicates

We simulated data for three source vectors $S = (1,1)$, $S = (1,2)$, and $S = (1,2,1)$. For $S = (1,1)$ and $S = (1,2)$, we varied the mean read depth of sample $d = 2$ while keeping the mean depth of sample $d = 1$ constant at 1x. We tested five different λ values for sample $d = 2$: 1, 2, 4, 6, and 12. For $S = (1,2,1)$, we varied the mean read depth of sample $d = 3$ while keeping the mean depth of samples $d = 1$ and $d = 2$ constant at 1x. We again tested five λ values for sample $d = 2$: 1, 2, 4, 6, and 12. We held L constant at 1000 across all experiments and tested the same five MAF intervals as before.

Comparing results to hierarchical clustering

To compare results from BIGRED and hierarchical clustering, we used genotyping-by-sequencing (GBS) data [8] collected by three of the four breeding programs collaborating on the NEXTGEN Project: the International Institute of Tropical Agriculture (IITA), the National Crops Resources Research Institute (NaCRRI), and the National Root Crops Research Institute (NRCRI). We refer the reader to the section "Data generation and variant calling" of [9] for a description of how the data were generated and filtered. We estimated non-replicate rates for

these three programs. The Euler diagram below shows the number of cases where a given genotype has $k > 1$ sequence records, for each breeding institution (Figure 2.4). We found $k = 9$ samples associated with TMEB419, a genotype used in breeding efforts at both IITA and NRCRI, and excluded this genotype from our analysis due to the computational demands for cases where $k > 7$. Appendix Figure 2.5 plots the number of source vectors associated with k for $k \in \{1, \ldots, 8\}$. We also removed putative replicates with a genome-wide mean read depth below 0.5. We ran BIGRED using $L = 1000$ randomly sampled sites across cassava's 18 chromosomes with MAFs falling between (0.4,0.5]. No two sites fell within 20 kb from one another, and we assumed a fixed sequencing error rate of 0.01 when calculating genotype likelihoods.

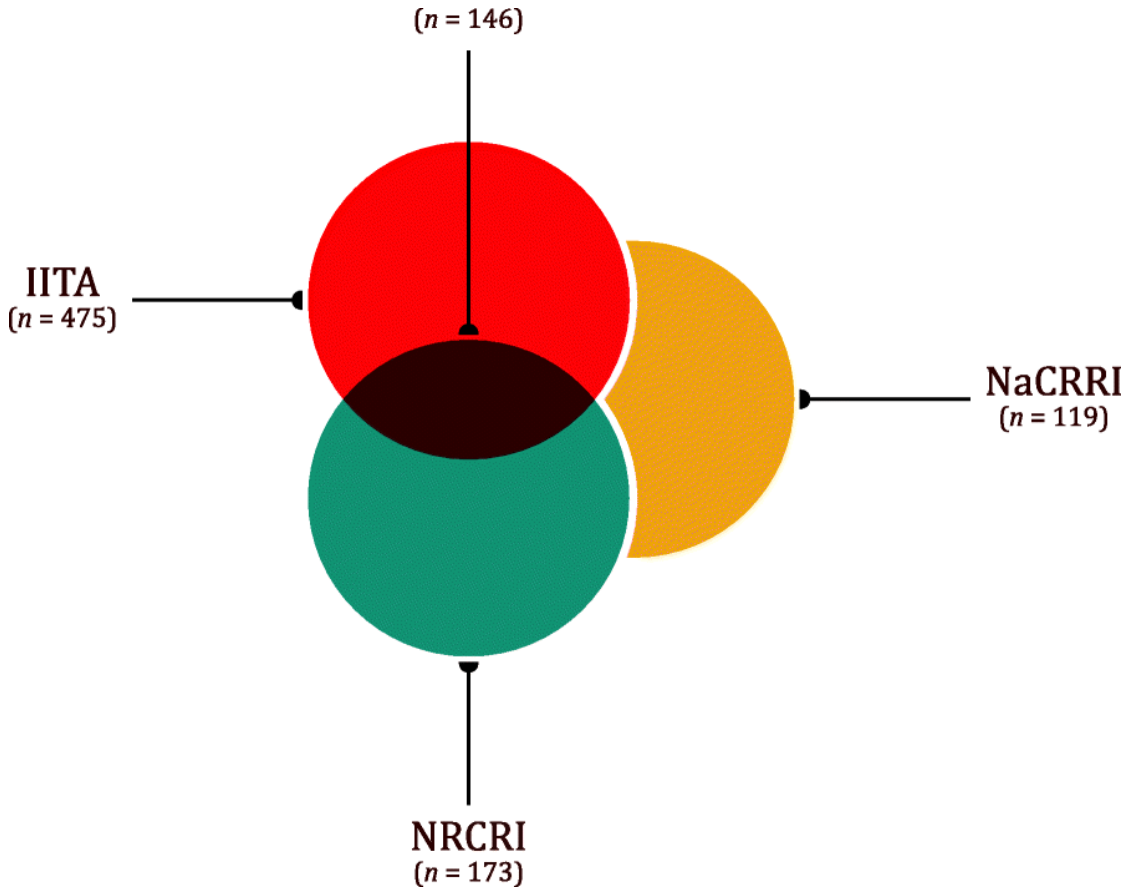


Figure 2.4 An Euler diagram showing the number of cases (*n*) where a given genotype has been sequenced more than once.
We found *n* = 475 genotypes (excluding TMEB419) within the IITA germplasm collection that have each been sequenced *k* >1 times. Entries falling at the intersection of IITA and NRCRI (black) represent cases where IITA submitted DNA for *k-x* sequence runs of a given genotype and NRCRI submitted DNA for the remaining *x* runs. There were 146 such cases. We found *n* = 173 genotypes within the NRCRI germplasm collection that have each been sequenced *k* >1 times. We found *n* = 119 genotypes within the NaCRRI germplasm collection that have each been sequenced *k* >1 times.

We compared results from BIGRED to results obtained from hierarchical cluster analysis. Results from [10] show that hierarchical clustering is an effective tool for matching accessions from farmers' fields to corresponding varieties in an existing database of known varieties, a problem very similar to the one being addressed in this paper. We performed hierarchical clustering on the *k* putative

replicates of each genotype. To do this, we first calculated the realized additive relationship matrix for the 1215 sequenced samples from IITA using sites harboring biallelic SNPs. Sites were filtered using criteria based on MAF and percent missing. Sites with a MAF falling within the interval (0.1,0.5] and with <50% missing data across the 1215 samples were kept, leaving us with 46,862 sites (out of 100,267) to analyze. We calculated the realized additive relationship matrix using the A.mat() function from the R package rrBLUP [11]. We used a matrix of genotype dosages as input and imputed missing dosage values using the "mean" option. We then calculated a distance matrix between the rows of the additive relationship matrix using Euclidean distance as the distance measure. We performed complete-linkage hierarchical clustering using the hclust() function and the distance matrix as input [12]. For each genotype, the hclust() function returns a tree structure with *k* leaves, each leaf representing a putative replicate. We determined the underlying relationship among each genotype's putative replicates by cutting each tree at a height of 0.5. We refer to this relationship as the "source vector" to keep terminology consistent with that of BIGRED's. We compared results from the complete-linkage cluster analysis to results from BIGRED. For BIGRED, we set a posterior probability threshold of 0.99, i.e., BIGRED would only return an inferred source vector if that source vector had a posterior probability of at least 0.99. This minimum posterior probability threshold was met in all cases, i.e., we were able to infer a source vector in all cases. We repeated this procedure for NaCRRI (299 sequenced samples and 48712 sites) and NRCRI (415 sequenced samples and 48320 sites).

For each breeding institution, we categorized the institution's genotypes into groups based on the number of putative replicates (k) each genotype had. We then calculated a mean non-replicate rate μ_k separately for each k. To calculate this, we computed a non-replicate rate for each individual that has k putative replicates (when $k = 2$, this rate is $1 - P(S=(1,1)|X)$), and then averaged these values across all individuals of a given k.

Comparing the consistency of BIGRED and hierarchical clustering

To compare the consistency of BIGRED and hierarchical clustering, we performed a set of experiments using the GBS data from the 475 IITA individuals with $1<k<7$ putative replicates. The basic premise of these experiments is that an analysis based on a larger set of sites is likely to be correct. The first step in these experiments is to perform error detection on an individual's putative replicates using the data at a large number of sites and to set the inferred source vector as the "truth". The second step is to perform error detection once more on the individual's replicates, this time using the data at a smaller number of sites disjoint from the initial set. To obtain a measure of consistency, we compare the results from the first (larger) analysis with results from the second (smaller) analysis.

To evaluate the consistency of hierarchical clustering, we first filtered the data, retaining samples with a genome-wide mean read depth of ≥0.5 and sites with MAFs within the interval (0.3,0.5] and with <50% missing data across the filtered samples. This left 1215 samples and 16,926 sites for analysis. As before, we called

genotype dosages using the observed allelic read depth data and imputed missing values at a given site with the site mean. We then performed hierarchical clustering on each of the 475 individuals, using data from 2000 randomly sampled sites. We set the output of these analyses as the "truth". We then performed hierarchical clustering on each of the individuals a second time, sampling *L* sites disjoint from the initial 2000, and compared the inferred source vector with the "true" source vector. We tested five values of *L*: 50, 100, 250, 500, and 1000. We repeated the experiment 10 times for each value of *L* and calculated a mean concordance rate between the "true" source vector and the source vector inferred from the *L* sites across the 10 runs and 475 cases for each *L*.

To evaluate the consistency of BIGRED, we first filtered the data, keeping samples with a genome-wide mean read depth of ≥0.5 and sites with MAFs within the interval (0.3,0.5]. As with hierarchical clustering, we defined the truth using 2000 randomly sampled sites. We used a fixed sequencing error rate of 0.01 and sampled sites such that no two sites fell within 20 kb from one another. We followed the same procedure as the one used to evaluate the consistency of hierarchical clustering, in particular, testing with the same five values of *L*.

Applying a pairwise-comparison approach to real data

Methods that employ a pairwise-comparison approach for error detection rather than joint analysis of the samples might produce ambiguous results when more than two putative replicates exist. To demonstrate, we applied a pairwise-comparison method to IITA's data, specifically we calculated the Pearson correlation

between all pairs of putative replicates. We refer to this method as the "correlation method". Before calculating the Pearson correlation between replicate pairs, we filtered the data, retaining samples with a genome-wide mean read depth of ≥0.5, sites with MAFs within the interval (0.3,0.5], and with <50% missing data across the filtered samples. This left 1215 samples and 16,926 sites for analysis. We called genotype dosages using the observed allelic read depth data and imputed missing values using glmnet [9]. We then calculated the Pearson correlation between all pairs of putative replicates using the cor() function [12]. For simplicity, we limited our analysis to the 154 cases where k=3. Correlations ranged from 0.02 to 0.93, so we selected 0.85 as the replicate-call threshold (i.e., two putative replicates with a correlation ≥0.85 are considered true replicates). We also applied a replicate-call threshold of 0.80 to examine how results changed.

Run time

We measured computation time as the number of central processing unit (CPU) seconds required to run BIGRED. All jobs were submitted to the Computational Biology Service Unit at Cornell University, which uses a 112 core Linux (CentOS 7.4) RB HPC/SM Xeon E7 4800 2U with 512GB RAM.

RESULTS

Evaluating the accuracy and run-time of BIGRED

To evaluate the algorithm's accuracy and run-time, we performed a full factorial experiment where we simulated data for each of the source vectors

associated with k = 2, 3, and 4, varying the mean read depth of samples and the MAF of the L = 1000 sites sampled by the algorithm. We use the term "accuracy" to refer to the median posterior probability of the true source vector. For these experiments, we simulated the situation where all k putative replicates had identical mean read depths but later tested the scenario where mean read depths varied among the k samples (refer to the section "Evaluating BIGRED's accuracy when mean read depths vary among the k putative replicates"). We observed qualitatively similar results for k = 2, 3, and 4, so we present only the results for k = 3 in the main text (Figure 2.5). We present the results for k = 2 and 4 in Appendix Figure 2.2 and 2.3, respectively. When no erroneous samples were present among the k putative replicates, the algorithm performed consistently well across all mean read depths and MAF intervals, assigning a median posterior probability of one to the true source vector (Figure 2.5A). We observed a common trend for the remaining two source vectors: for a given MAF interval, accuracy monotonically increased as mean read depth increased. We observed this trend in all cases except for interval (0.0,0.1], whose median accuracy stayed constant at zero across all depths for S = (1,2,1) and S = (1,2,3) and intervals (0.3,0.4] and (0.4,0.5], whose median accuracies stayed constant at one across all depths for S = (1,2,1) and S = (1,2,3) (Figure 2.5B and 5C). In addition to recording the posterior probability of the true (simulated) source vector, we also recorded the posterior probability assigned to all other source vectors. We present the plots for S = (1,2,1) and S = (1,2,3) experiments in Appendix Figure 2.4. These plots recapitulate the behavior observed in Figure 2.5 but do so at a higher resolution: for a given MAF interval, with the exception of

(0.0,0.1], BIGRED shifts the probability away from S=(1,1,1) towards the true (simulated) source vector as the mean read depth of samples increases. The algorithm takes, on average, approximately three seconds to analyze all possible source vectors when the true source vector is S = (1,1,1) for all pairwise combinations of sample mean read depth and site MAF interval (Figure 2.5D). Similarly, the algorithm takes, on average, approximately four seconds to analyze all possible source vectors when the true source vectors were S = (1,2,1) and S = (1,2,3) for all pairwise combinations of sample mean read depth and site MAF interval (Figure 2.5E and 5F).

To assess the impact of L on the algorithm's accuracy, we repeated simulation experiments for S = (1,2,1) and S = (1,2,3), this time varying values of L and looking only at sites with MAFs falling in (0.2,0.3]. We tested the (0.2,0.3] interval since median accuracy was one for all earlier experiments using intervals (0.3,0.4] and (0.4,0.5]. We tested seven values of L: 50, 100, 250, 500, 1000, 2000, and 5000. Median accuracy drastically increased when L increased from 100 to 250 for S = (1,2,1) at 2x mean depth (Figure 2.6A). At a given mean read depth, we observed little to no change in median accuracy when increasing L for S = (1,2,3) (Figure 2.6B).

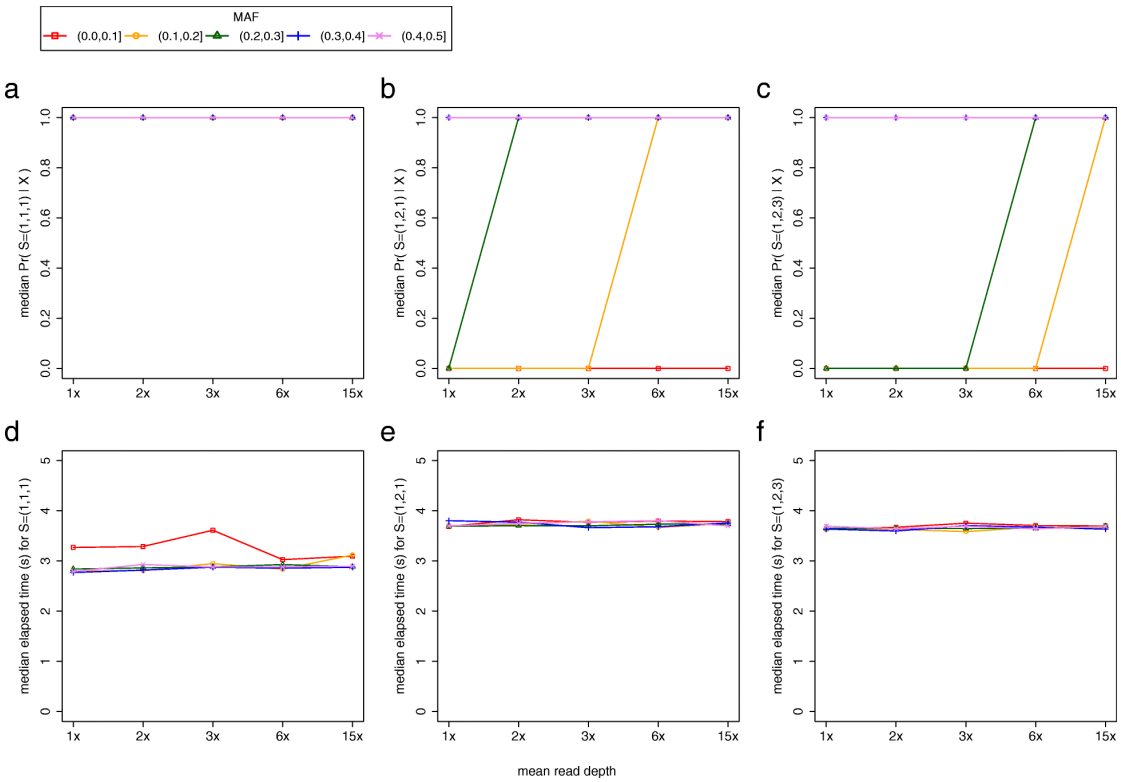


Figure 2.5 Algorithm's accuracy and run-time as a function of the mean read depth of samples and the MAF of analyzed sites for $k = 3$.
(A, B, and C) Each plot shows estimates of the median posterior probability of the true source vector (y-axis) as a function of mean read depth of samples (x-axis) and MAF of sites (legend). Each data point presents the median posterior probability of $S = (1,1,1)$ across 15 runs, $S = (1,2,1)$ across 100 runs, and $S = (1,2,3)$ across 100 runs of the algorithm. (D, E, and F) Each plot shows the mean elapsed time in seconds for each simulation scenario.

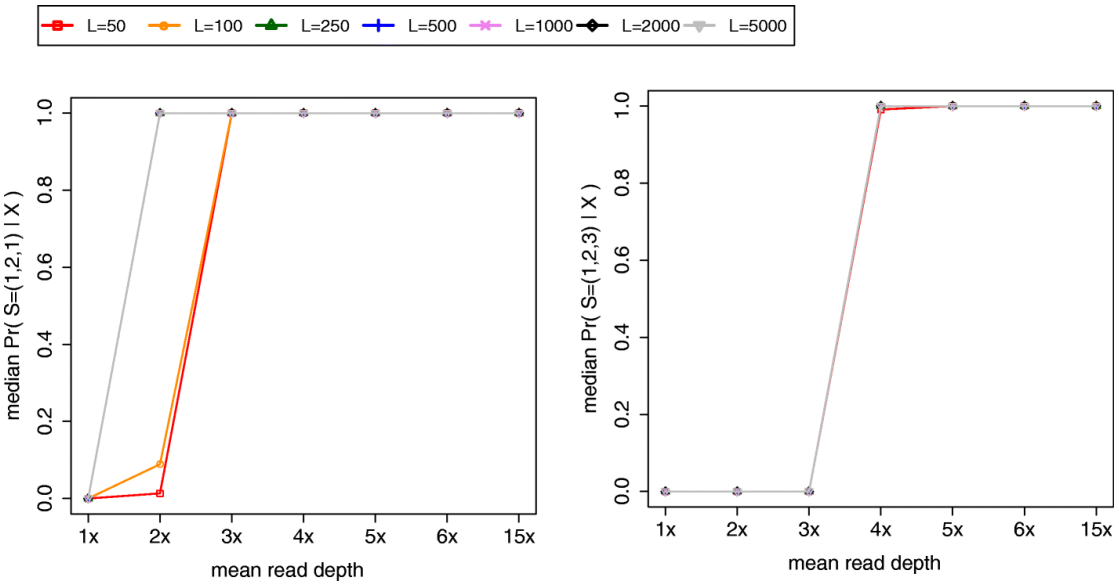


Figure 2.6 The impact of *L* on accuracy.
The two plots show estimates of the median posterior probability of the true source vector (*y*-axis) as a function of mean read depth of samples (*x*-axis) for different values of *L* (legend). We sampled sites whose MAFs fell in the interval (0.2,0.3].

Evaluating the sensitivity of the algorithm:

To evaluate the algorithm's sensitivity, we first simulated the scenario where S = (1,1) then contaminated p percent of sites in sample d = 2 with a second genotypic source. We then assessed how much probability the algorithm assigned to source vector S = (1,1) in light of these contaminated sites. We tested five different values of p in combination with five sample mean read depths. The algorithm showed greater sensitivity to increases in p as the mean read depth of the samples increased (Figure 2.7).

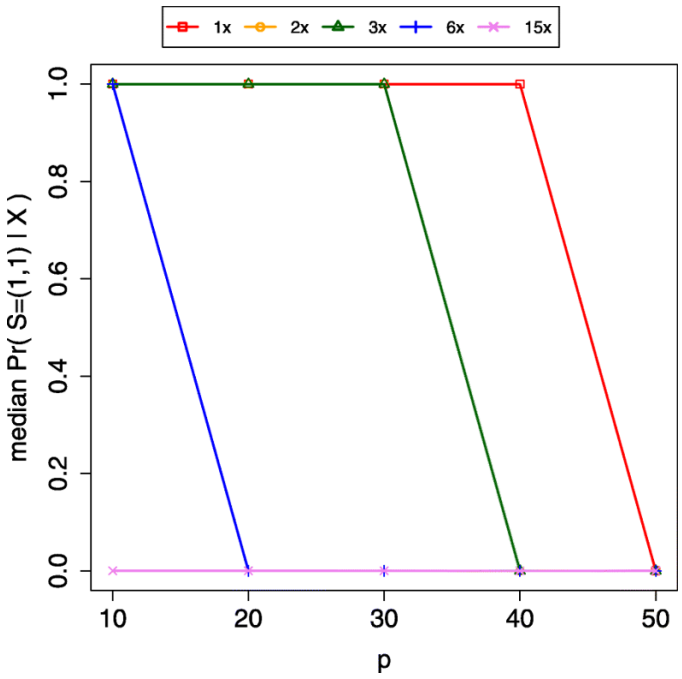


Figure 2.7 Algorithm's sensitivity as a function of the mean read depth of samples.
We assessed the impact of mean read depth on the method's sensitivity. The plot reports estimates of the median posterior probability of the true source vector S = (1,1) (y-axis) as a function of the percentage of contaminated sites (p) in sample d = 2 (x-axis) and mean read depth of putative replicates (legend). In these experiments, samples d = 1 and d = 2 have identical mean read depths.

Evaluating BIGRED's accuracy when mean read depths vary among the k putative replicates

We next evaluated the algorithm's accuracy when the read depths vary among the k samples. For these experiments, we examined three source vectors S = (1,1), S = (1,2), and S = (1,2,1) and used L = 1000 sites. And as before, we examined the impact of MAF at the 1000 sites. When simulating data for source vectors S = (1,1) and S = (1,2), we varied the mean read depth of sample d = 2 while keeping the mean depth of sample d = 1 constant at 1x. We tested five different read depth values for sample d = 2 (λ = 1, 2, 4, 6, and 12). When simulating data for source vector S = (1,2,1), we varied the mean read depth of sample d = 3 while keeping the

mean depth of samples d = 1 and d = 2 constant at 1x. We tested five different read depth values for sample d = 3 (λ = 1, 2, 4, 6, and 12). We obtained results comparable to those from simulation experiments where all k putative replicates had identical mean read depths. For S = (1,1), the algorithm performed consistently well across all read depth differences and MAF intervals, assigning a median posterior probability of one to the true source vector (Figure 2.8A). For S = (1,2) and S = (1,2,1), the algorithm performed consistently well across all read depth differences when analyzing sites with MAFs falling in (0.3,0.5] and consistently poorly across all read depth differences when analyzing sites with MAFs falling in (0.0,0.2] (Figure 2.8B and 8C). For MAF interval (0.2,0.3], median accuracy monotonically increased as the difference between sample read depths grew, i.e. as the mean read depth for sample d = 2 in S = (1,2) and d = 3 in S = (1,2,1) increased (Figure 2.8B and 8C).

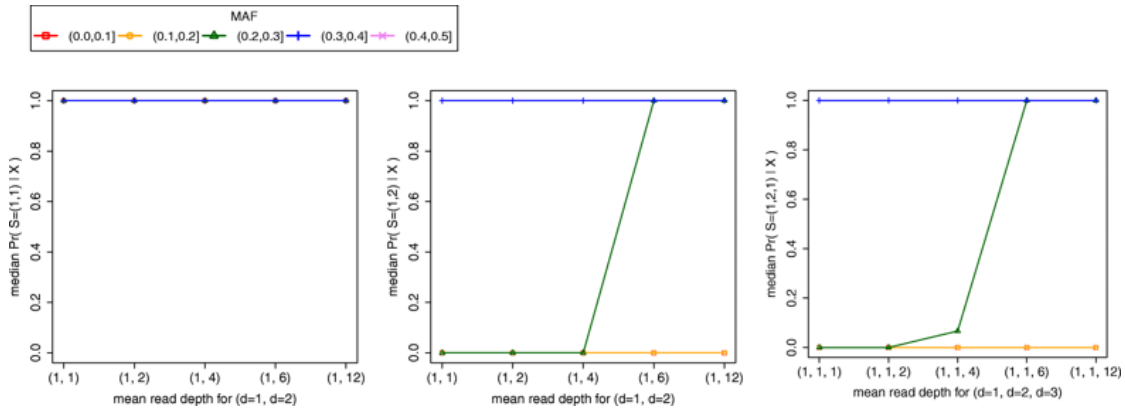


Figure 2.8 Accuracy of the algorithm when the mean read depths of the *k* putative replicates vary

Each data point in the three plots reports the median posterior probability for the true source vector (*y*-axis) as a function of the mean read depth for the *k* samples (*x*-axis) and the MAF of sampled sites (legend).

Estimating NEXTGEN non-replicate rates

We estimated non-replicate rates μ_k for IITA, NaCRRI, NRCRI, and the germplasm used by both IITA and NRCRI, respectively (Table 2.1).

Table 2.1 A table summarizing the mean non-replicate rate μ_k of each breeding institution.
For each institution, we categorized genotypes into groups based on the number of putative replicates each genotype had. Grey rows show the number of genotypes in each group n_k for each breeding institution. We then calculated the mean non-replicate rate among genotypes of a given k μ_k by calculating the mean probably of no errors then subtracting this value from one.

	Institution	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
n_k	IITA	272	154	37	11	1
μ_k	IITA	0.21	0.16	0.14	0.27	1
n_k	NaCRRI	58	61	0	0	0
μ_k	NaCRRI	0.05	0.21	-	-	-
n_k	NRCRI	128	31	5	8	1
μ_k	NRCRI	0.37	0.32	0.40	0.25	1
n_k	IITA & NRCRI	101	31	5	8	1
μ_k	IITA & NRCRI	0.33	0.32	0.40	0.25	1

Method comparison

We compared results from BIGRED to results obtained from complete-linkage hierarchical cluster analysis. The two methods reported 28, 2, and 15 conflicting results for IITA, NaCRRI, and NRCRI, respectively (Figure 2.9), all of which were cases where hierarchical clustering reported an error among putative

replicates while BIGRED reported no error, with the exception of one NRCRI individual UG120041. Both methods reported an error for UG120041 but reported different errors: BIGRED inferred a (1,2,3) relationship while hierarchical clustering inferred a (1,1,2) relationship.

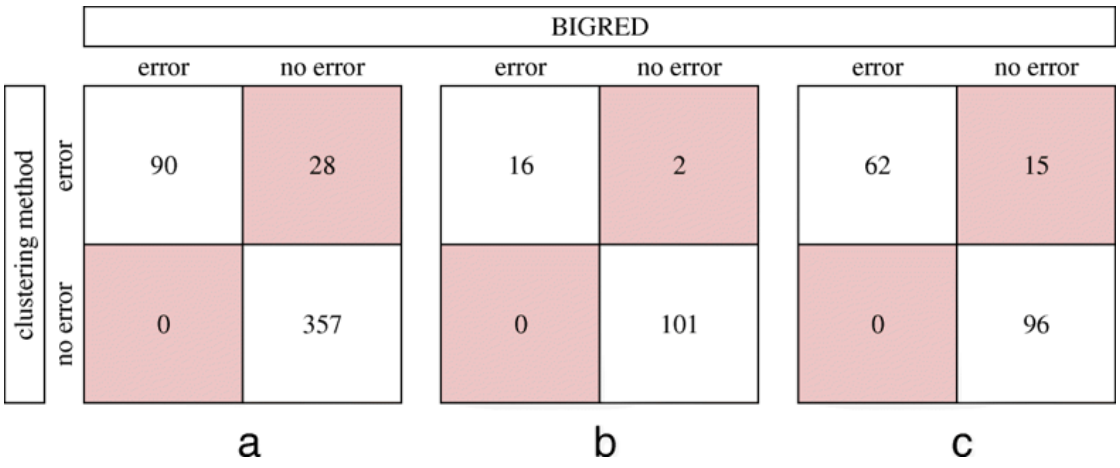


Figure 2.9 Comparing results from complete-linkage hierarchical clustering and the proposed method
Above are three two-way contingency tables comparing the results from complete-linkage hierarchical cluster analysis and the proposed method for IITA (A), NaCRRI (B), and NRCRI (C). Conflicts between the two methods are shown in red. The 146 genotypes shared between IITA and NRCRI (Figure 2.4; black) are represented twice in our results: once with the 329 genotypes unique to IITA and once with the 27 genotypes unique to NRCRI.

We compared the consistency of BIGRED with that of hierarchical clustering. Table 2.2 presents the mean concordance rate between the "true" source vector and the source vector inferred from *L* sites among 475 cases across the 10 runs of hierarchical clustering and BIGRED. BIGRED had a higher concordance rate than hierarchical clustering at every *L*, suggesting that BIGRED is a more consistent estimator than hierarchical clustering.

Table 2.2 A table comparing the consistency of BIGRED and hierarchical clustering using the 475 IITA individuals with 1<*k*<7 putative replicates.
To evaluate the consistency of the two methods, we performed error detection on an individual's putative replicates using the data at 2000 sites and set the inferred source vector as the "truth". We then performed error detection a second time using a smaller number of sites (*L*) disjoint from the initial set. We compared the "true" source vector with the source vector inferred from *L* sites. For each IITA individual, we tested five values of *L* and repeated the experiment 10 times for each value of *L*. We then calculated the mean concordance rate between the "true" source vector and the source vector inferred from *L* sites across the 475 cases and across 10 runs.

Method	*L*=50	*L*=100	*L*=250	*L*=500	*L*=1000
BIGRED	0.9832	0.9895	0.9958	0.9973	0.9981
Hierarchical clustering	0.8322	0.9088	0.9488	0.9640	0.9771

One motivation for BIGRED's joint analysis framework is that pairwise-comparison methods might produce ambiguous results for cases of more than two putative replicates. We introduced a hypothetical example of this in the Background section and found real examples of these inconsistencies when applying a pairwise-comparison method to IITA's data. More specifically, when examining cases of *k*=3 and using a replicate-call threshold of 0.85, we found 80 cases (out of 154) where the pairwise method awarded any pair of samples (of an individual) replicate status. Of these 80 cases, we found 10 cases where the method produced ambiguous results. When we decreased the call threshold to 0.80, we found 146 cases where the method inferred at least one true replicate pair but six of these cases had ambiguous results.

DISCUSSION

Researchers may choose, for a number of reasons, to sequence a given individual more than once. Regardless of intent, it is important to identify potentially mislabeled or contaminated samples before using the data (e.g. merging the data from replicate sequence runs or using the data to optimize bioinformatics quality filters). Unfortunately, existing methods to detect such errors are ad hoc and ill suited for use in shallow-depth HTS data since they require some combination of genotype calling, imputation, and haplotype phasing. We have introduced a new probabilistic framework for error detection that addresses key limitations of existing methods. Using Bayes Theorem, we calculate the posterior probability distribution over the set of relations describing the putative replicates (i.e. the set of source vectors), allowing us to infer which of the samples originated from an identical genotypic source.

We examined the impact of mean read depth, L, and MAF at the L sites on the accuracy of the proposed method through a series of simulation experiments. We found that the algorithm is most accurate when analyzing sites whose MAFs fall in the range (0.3,0.5], consistently across all mean read depths when $L = 1000$ (Figure 2.5). Sites with MAFs falling in the interval (0.0,0.1] relay little information to the algorithm. When analyzing these sites, BIGRED assigns a median posterior probability of one to $S = (1,1,1)$, regardless of the true source vector. Thus BIGRED appears to be biased towards inferring no error among putative replicates when analyzing sites with low MAF. One reason for this bias is our definition of $P(G^{(v)}|S)$ (Figure 2.2). Given a site that has a reference allele frequency of 0.1, when $k = 3$, the probability of $G^{(v)} = (AA,AA,AA)$ given $S = (1,1,1)$, i.e. no erroneous samples among

the putative replicates, is 0.1^2, whereas the probability of $G^{(v)}$ = (AA,AA,AA) given any other source vector is ≤ 0.1^4. This bias is compounded by the fact that we estimated allele frequencies from a set of 206 individuals but ran simulation experiments using a subset of 15. Some loci that had low but non-zero MAF among the 206 individuals appeared monomorphic among the 15 individuals, making the 15 individuals look more similar than they actually are in reality. We found that 47.14% and 5.29% of sites with MAFs in the (0.0,0.1] and (0.1,0.2] interval, respectively, became monomorphic among the 15 individuals.

To evaluate the impact of L on the algorithm's accuracy, we repeated simulation experiments for S = (1,2,1) and S = (1,2,3) using different values of L and looking only at sites with MAFs falling in (0.2,0.3]. Surprisingly, we observed little to no change in median accuracy at a given depth when increasing the number of sampled sites. The only exception was S = (1,2,1) at 2x mean depth, where we observed a drastic increase in accuracy when increasing L from 100 to 250 (Figure 2.6). For S = (1,2,3) at 2x and 3x, we observed a median accuracy of zero even when sampling 5000 sites. We observed an increase in median accuracy only after increasing the mean read depth of samples to 4x. These results indicate that the mean read depth of samples contributes more to accuracy than the number of sampled sites. In these simulation experiments, all k putative replicates of a given genotype were assigned identical mean read depths. These results, however, were robust to samples with varying mean read depths (Figure 2.8).

We also assessed the sensitivity of the algorithm as a way to gauge how the proportion of exogenous DNA affects the algorithm and how allelic sampling bias

impacts results. The GBS protocol uses methylation-sensitive restriction enzymes (REs) to avoid sampling highly repetitive regions of the genome. One potential complication when using methylation-sensitive REs is allelic sampling bias of a marker or unequal sampling and sequencing of homologous chromosomes, resulting from differential methylation in a region. ApeKI, the RE employed by NEXTGEN, for instance, will not cut if the 3' base of the recognition sequence on both strands is 5-methylcytosine. To test the impact of imperfect marker "heritability", we simulated the scenario where $S = (1,1)$ and corrupted p percent of sites in sample $d = 2$ with a second genotype source. We tested the cases where $p = \{10\%,20\%,30\%,40\%,50\%\}$ for five different sample mean depths ($\lambda = \{1,2,3,6,15\}$) and found that the algorithm was robust to increases in p for lower values of λ (Figure 2.7). Not surprisingly, the method assigned higher probability to $S = (1,2)$ as p and mean depth increased. As mean depth increases, the algorithm grows increasingly confident that differences at sites reflect true biological differences rather than sampling variation or error.

When applying BIGRED and hierarchical clustering on real data, we found a relatively high concordance rate between the two methods (Figure 2.9). Although this comparison does not directly tell the reader which of the two methods is more accurate, the comparison and the analyses in this paper demonstrate the benefits of using BIGRED over hierarchical clustering. Firstly, we found that BIGRED is a more consistent estimator relative to hierarchical clustering (Table 2.2). Secondly, BIGRED employs a probabilistic framework to tackle the problem of error detection rather than a heuristic one like hierarchical clustering, making BIGRED a more

statistically rigorous and neatly packaged method. Hierarchical clustering requires the user to make many (arguably arbitrary) decisions throughout the protocol, whereas BIGRED requires the user to make one decision at the very end, i.e. the probability at which to "call" a source vector. Our results also highlight one of the major flaws of methods like hierarchical clustering: results can change depending on what samples were included in the analysis, specifically during imputation. There are 146 genotypes that are used in both IITA's and NRCRI's breeding programs, and these 146 genotypes appear in both institutions' data (Figure 2.4). We performed hierarchical clustering on these individuals a total of two different times: once in combination with the 329 genotypes unique to IITA and once in combination with the 27 genotypes unique to NRCRI. Ideally, the duplicate runs of an individual would produce identical results, regardless of what other samples where included in each analysis. Of the 146 cases, however, we found three cases where the hierarchical clustering-based duplicate analyses produced conflicting results: one case where the two analyses reported differ errors and two cases where the IITA analysis reported no error but the NRCRI analysis reported an error. These conflicts likely resulted from the imputation component of the cluster analysis procedure since sample composition is known to affect imputation. These issues highlight the benefits of our approach: when we ran BIGRED on these 146 individuals twice, we found that all duplicate runs produced identical results.

In our simulation experiments, we estimated allele frequencies from WGS data. Users of BIGRED will likely not have this option and will need to estimate allele frequencies using low- to moderate-depth sequence data. Although such frequency

estimates will in general contain noise, we showed that BIGRED is robust to imperfect estimates of allele frequency. We estimated allele frequencies from a set of 206 individuals but ran simulation experiments using a subset of 15 individuals and were able to recover the true underlying source vector when analyzing sites with MAFs falling in the (0.3,0.5] interval (Figure 2.5). We also suggest that a user perform preliminary analyses (e.g., with PCA) to detect the presence of population structure, and when structure is evident, we recommend analyzing subpopulations separately, estimating allele frequencies from samples of a given subpopulation then running BIGRED on the samples from that subpopulation.

The number of possible source vectors increases exponentially as k increases (Appendix Figure 2.5). For this reason, we do not recommend using BIGRED on cases where $k>7$. We, however, do not anticipate many scenarios where a researcher would have sequenced a given individual more than seven times, but if this scenario does occur, one could either randomly select seven putative replicates to analyze or divide the replicates into sets of no more than seven samples. If using the latter scheme, one would run BIGRED on each set, merge the true replicates within each set (discarding the erroneous samples), then combine the sets of merged samples before running BIGRED once more. By using a Poisson distribution to simulate AD data, we make the assumption that reads are uniformly distributed across the genome. While read data will often be more highly dispersed than these analyses, if at least L of the sites in those data have read depth of lambda or above, BIGRED will perform at least as well as in our analyses with these same parameters.

A motivation for BIGRED's joint analysis framework is that pairwise-comparison methods might produce ambiguous results when more than two putative replicates exist, and we did, in fact, run into cases of this when applying the correlation method to real data. Of the cases where the method reported the presence of replicates when applying a replicate-call threshold of 0.85 and 0.80, 12.50% and 4.11% contained pairwise inconsistencies, respectively. By decreasing the call threshold, one lowers the number of ambiguous cases returned but doing so also increases the number of false positives returned. And although it may occur at low frequency, the possibility of pairwise inconsistencies exists and would be a problem for all methods that employ a pairwise-comparison approach.

In this study, we introduced a statistical framework for detecting mislabeled and contaminated samples among putative replicates. Our method addresses key limitations of existing approaches and produced highly accurate results in simulation experiments even when applied to samples with low read depth. Our method is implemented as an R package called BIGRED, which is freely available for download: https://github.com/ac2278/BIGRED.

REFERENCES

[1] K. Robasky, N. E. Lewis, and G. M. Church, "The role of replicates for error mitigation in next-generation sequencing," *Nature Reviews Genetics*, vol. 15, no. 1. pp. 56–62, 2014.

[2] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: A Tool

Set for Whole-Genome Association and Population-Based Linkage Analyses," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, 2007.

[3] R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song, "Genotype and SNP calling from next-generation sequencing data.," *Nat. Rev. Genet.*, vol. 12, no. 6, pp. 443–51, 2011.

[4] "NEXTGEN Cassava." [Online]. Available: http://www.nextgencassava.org/.

[5] P. Ramu, W. Esuma, R. Kawuki, I. Y. Rabbi, C. Egesi, J. V. Bredeson, R. S. Bart, J. Verma, E. S. Buckler, and F. Lu, "Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation," *Nat. Genet.*, 2017.

[6] X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir, "A high-performance computing toolset for relatedness and principal component analysis of SNP data," *Bioinformatics*, 2012.

[7] J. V Bredeson, J. B. Lyons, S. E. Prochnik, G. A. Wu, C. M. Ha, E. Edsinger-Gonzales, J. Grimwood, J. Schmutz, I. Y. Rabbi, C. Egesi, P. Nauluvula, V. Lebot, J. Ndunguru, G. Mkamilo, R. S. Bart, T. L. Setter, R. M. Gleadow, P. Kulakow, M. E. Ferguson, S. Rounsley, and D. S. Rokhsar, "Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity," *Nat Biotech*, vol. advance on, Apr. 2016.

[8] R. J. Elshire, J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell, "A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species," *PLoS One*, vol. 6, no. 5, 2011.

[9] A. W. Chan, M. T. Hamblin, and J.-L. Jannink, "Evaluating Imputation Algorithms for Low-Depth Genotyping-By-Sequencing (GBS) Data.," *PLoS One*,

vol. 11, no. 8, p. e0160733, 2016.

[10] I. Y. Rabbi, P. A. Kulakow, J. A. Manu-Aduening, A. A. Dankyi, J. Y. Asibuo, E. Y. Parkes, T. Abdoulaye, G. Girma, M. A. Gedil, P. Ramu, B. Reyes, and M. K. Maredia, "Tracking crop varieties using genotyping-by-sequencing markers: A case study using cassava (Manihot esculenta Crantz)," *BMC Genet.*, 2015.

[11] J. B. Endelman, "Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP," *Plant Genome J.*, 2011.

[12] R Development Core Team, "R: A Language and Environment for Statistical Computing," *R Found. Stat. Comput. Vienna Austria*, 2016.

CHAPTER 3

CHARACTERIZING RECOMBINATION IN MANIHOT ESCULENTA

ABSTRACT

Recombination has essential functions in evolution, meiosis, and breeding. Here, we use the multi-generational pedigree and genotyping-by-sequencing (GBS) data from the International Institute of Tropical Agriculture (IITA) to characterize recombination in cassava (*Manihot esculenta*). We detected recombination events using SHAPEIT2 and duoHMM, characterized the recombination landscape across the 18 chromosomes of cassava, constructed a genetic map and compared it to an existing map constructed by the International Cassava Genetic Map Consortium (ICGMC), and constructed sex-specific genetic maps to see if there's evidence of sexual dimorphism in crossover distribution and frequency. The IITA pedigree consists of 7,165 informative meioses (3,122 female; 3,099 male).

INTRODUCTION

Although mutations introduce new genetic variation in a population, the most important mechanism for generating genomic diversity in sexually reproducing species is the production of new combinations of already existing alleles, or recombination, a process that occurs during prophase I of meiosis through crossing-over. Meiotic recombination increases the probability that offspring from two individuals will carry combinations of alleles that allow survival and reproduction in a changing environment. In the context of plant breeding,

recombination is important because it dictates the resolving power of quantitative trait mapping and the precision of allele introgression. Recombination has an additional essential function in that it aids in homology recognition and helps ensure proper disjunction, or segregation of homologous chromosomes during meiosis [1]. Improper disjunction, or nondisjunction, results in aneuploidy, a deleterious outcome in which gametes have more or less than the typical chromosome number.

The number of crossovers per chromosome and the distribution of crossovers along chromosomes are tightly controlled. Crossover number appears to be constrained by both an upper and lower bound. The reason for a lower bound on crossover number is clear since in most species, there is a need for one obligatory crossover per chromosome pair to prevent aneuploidy. The reason(s) for an upper bound on crossover number, however, is less obvious. Results from a recent study, where crossover rate was significantly increased in mutant *Arabidopsis thaliana*, suggest that reduced fertility (in the form of reduced pollen viability and seed set) may be associated with increased recombination [2]. One plausible evolutionary explanation for the existence of an upper bound on recombination is that beneficial alleles residing on the same haplotype may collectively act to increase fitness. Recombination can break these associations, resulting in reduced progeny fitness [3].

The distribution of crossovers along chromosomes is not random and is influenced by chromosome features such as chromatin structure, gene density, and nucleotide composition. Chromatin structure strongly influences the position of

crossovers along chromosomes. Like in other eukaryotes, crossovers in plants occur more frequently in decondensed, euchromatic regions and less frequently in highly condensed, heterochromatin regions. The occurrence of a crossover at one location also reduces the likelihood that another crossover will occur in close proximity. This nonrandom placement of crossovers, known as chromosomal interference, results in a pattern where recombination events appear evenly spaced [4]. If only a limited number of crossovers can occur per meiosis, interference will result in crossovers being more evenly distributed across chromosomes (interference lowers the number of crossovers on large chromosomes and the remaining possible crossovers are more likely to occur on small chromosomes). Interference may therefore serve as a biological mechanism to ensure that every pair of homologous chromosomes undergoes at least one crossover event, which is necessary for proper disjunction.

In many species, crossover frequency and distribution along chromosomes differs between female and male meiosis, a phenomenon referred to as heterochiasmy [5]. The direction and degree of these differences are species-specific, and most extreme are cases in which one of the two sexes lacks meiotic recombination entirely. Male *Drosophila melanogaster*, for example, do not recombine during meiosis. To date, no investigation of sexual dimorphism has been conducted in cassava.

Cassava is a diploid organism with an estimated genome size of approximately 772 Mb spread across 18 chromosomes with the reference genome spanning 582.28 Mb [6]. The International Cassava Genetic Map Consortium (ICGMC) generated a consensus genetic map of cassava that combines 10 mapping

populations [7]. The 10 mapping populations consisted of one self-pollinated cross and nine biparental crosses (14 parents total; 3,480 meioses). The genetic map is 2,412 cM in length and organizes 22,403 GBS markers on 18 chromosomes. Here, we used the multi-generational pedigree from the International Institute of Tropical Agriculture (IITA) to characterize recombination in cassava. We used duoHMM-corrected, SHAPEIT2-inferred haplotypes to detect SNP intervals flanking a crossover event then used these intervals to map the recombination landscape across cassava's 18 chromosomes [8]. We built a genetic map from 7,165 meioses, compared it to ICGMC's composite map, and constructed sex-specific genetic maps to see if crossover distribution and frequency differ significantly between the two sexes. We also examined if there's evidence of chromosomal interference.

MATERIALS AND METHODS

A description of the IITA germplasm population structure

The IITA pedigree consists of 7,432 unique individuals. Each individual belongs to one of four breeding groups: Genetic Gain (GG; n = 494), TMS13 (n = 2,334), TMS14 (n = 2,515), or TMS15 (n = 2,089). Of the 494 GG individuals, 258 are the progeny of GG-GG crosses and the remaining 236 individuals are founders (individuals with no parents). All TMS13 members are the progeny of GG-GG crosses. Of the 2515 TMS14 individuals, 1,881 are the progeny of TMS13-TMS13 crosses and the remaining are GG-GG progeny. The TMS15 groups consists of 920, seven, 1,159, and three individuals that are progeny of TMS14-TMS14, TMS13-TMS14, TMS13-TMS13, and TMS13-GG crosses, respectively.

Merging replicate GBS records of each proband

We found GBS data for 7,294 of the 7,432 IITA individuals (n_{GG} = 366, n_{TMS13} = 2330, n_{TMS14} = 2509, and n_{TMS15} = 2089). Of the 366 GG individuals, 189 had more than one GBS record (i.e., NEXTGEN sequenced these 189 individuals multiple times). We refer to multiple sequence records of an individual as "replicates". Before merging the data from replicate sequence runs of an individual, we verified that no erroneous samples existed among the putative replicates (i.e. verified that all putative replicates derived from an identical individual). We validated putative replicates of an individual using BIGRED [9]. Using Bayes Theorem, BIGRED calculates the posterior probability distribution over the set of relations (i.e., source vectors) describing the putative replicates of an individual and infers which of the samples originated from an identical genotypic source. Of the 189 GG BIGRED runs, 21 produced ambiguous results. An ambiguous BIGRED result occurs when BIGRED returns a source vector where no source has a clear majority (e.g. S = (1,2) is the ambiguous source vector for the case where an individual has two putative replicates and S = {(1,1,2,2), (1,2,2,1), (1,2,1,2)} for the case where an individual has four putative replicates). Because we were unable to resolve these cases, we excluded these 21 GG individuals from future analyses. We merged the data for the 168 GG individuals with unambiguous BIGRED results, merging only the samples that were inferred to be true replicates. We repeated this process for TMS13 and TMS14 individuals (all individuals in the TMS15 group were sequenced once). Of the 2,330 TMS13 individuals, 156 had more than one GBS record and 10 produced

ambiguous BIGRED results. Of the 2,509 TMS14 individuals, 62 had more than one GBS record, and of the 62 BIGRED runs, three produced ambiguous results. We excluded these 13 TMS13 and TMS14 individuals from further analyses. Table 1 summarizes these results.

Table 3.1 Summary of data records for each breeding group.
The table shows the number of individuals, the number of individuals that were sequenced more than once, and the number of individuals with ambiguous BIGRED results for each breeding group.

Group ID	Number of individuals in group	Number of individuals with >1 sequence record	Number of individuals with ambiguous BIGRED results
GG	366	189	21
TMS13	2330	156	10
TMS14	2509	62	3
TMS15	2089	0	0

Validating IITA pedigree records using AlphaAssign

Of the remaining 345 (=366-21) GG individuals listed in the pedigree, 187 GG individuals had at least one listed parent with locatable GBS data. These parents also belong to the GG population. We used the parentage assignment algorithm AlphaAssign to validate the existing pedigree information for these 187 GG individuals [10]. AlphaAssign frames the parentage assignment problem as a relationship classification problem. Rather than directly attempting to identify target individual *t*'s parent from a list of candidate individuals, AlphaAssign, instead, attempts to classify the relationship between target individual *t* and each candidate individual *c*. AlphaAssign considers four possible target-candidate relationships, *H*:

the candidate individual is (1) a parent of the target individual, (2) a full-sibling of the target individual's parent, i.e., the target individual's uncle, (3) a half-sibling of the target individual's parent, and (4) unrelated to the target individual. Specifically, AlphaAssign calculates the posterior probability of these four relations, given the observed allelic depth (AD) data of *c* and *t* (and if known, the AD data for a known parent of *t*; if individual *t* has no known parent, the algorithm makes use of a 'dummy parent' whose genotype probabilities at a given site are calculated using estimated allele frequencies and assuming Hardy-Weinberg Equilibrium).

Informally, the algorithm first calculates for each biallelic site, the posterior genotype probabilities for target *t* given the observed AD data for target *t* and the posterior genotype probabilities for target *t*'s known or dummy parent. It then uses these two posterior genotype probability distributions to generate four 'proposal distributions' for candidate individual *c*, one proposal distribution for each of the four possible ways *c* and *t* are related. The proposal distribution for relationship h = 1, for example, gives the genotype probabilities for individual *c* given that *c* is a parent of target *t*. One can think of these proposal distributions as genotype priors. AlphaAssign then calculates the posterior probability of h = 1 by combining the genotype likelihood for candidate *c*, the proposal distribution given h = 1, and the prior distribution for H across all sites (AlphaAssign assumes that all sites segregate independently and a uniform prior on H) and dividing by some normalizing constant. Once posterior probabilities for H are calculated, AlphaAssign assigns each candidate an assignment score:

$$score = -\log\left(1 - \Pr(h = 1 | X_c, X_t, X_k)\right) \quad (1)$$

for the case where we know one parent of t

$$score = -\log(1 - \Pr(h = 1 | X_c, X_t, p_A)$$
for the case where we know neither parent

, where X_c, X_t and X_k represent the AD data for candidate *c*, target *t*, and a known parent of target *t*, respectively. Again, if target *t* has zero known parents, AlphaAssign makes use of a 'dummy parent' whose genotype probabilities at a given site are calculated using the estimated allele frequency of the reference allele p_A and assuming HWE. For AlphaAssign to assign candidate *c* as target *t*'s parent, candidate *c* must pass three criteria: (1) *c* must be classified as a parent (i.e. $h = 1$ must have the highest posterior probability), *c*'s assignment score must pass a threshold, and (3) *c*'s assignment score must be the highest among all candidates.

Because AlphaAssign looks at the relationship between pairs of individuals rather than among triplets, we ran AlphaAssign a total of two times to validate IITA's pedigree information. We walk through the validation procedure for GG individuals. In the first run, we provided the algorithm with no pedigree information (i.e., all calculations involved the use of a dummy parent). For each target individual, we listed all GG individuals as candidate parents (we did not list an individual as it's own candidate parent). We fed the algorithm AD data from 1,000 randomly sampled sites across cassava's 18 chromosomes. We sampled sites such that no two sites fell within 20 kb from one another. For each target individual, we identified the candidate individual with the highest score statistic and listed this top-scoring candidate as the target individual's parent in a (newly created) pedigree file. We ran AlphaAssign a second time, this time providing AlphaAssign with pedigree information, i.e., the AlphaAssign-inferred pedigree generated from the results of

the first run. We again identified the candidate individual with the highest score statistic for each target individual. Upon completing the two runs, each target individual had two AlphaAssign-inferred parents. We compared the AlphaAssign-inferred pedigree with IITA's existing pedigree. We repeated this analysis for the TMS13, TMS14, and TMS15 group and present the results of all four breeding groups in Table 2. We built the list of candidate individuals for each breeding group based on how IITA generated each group (refer to the section "*A description of the IITA germplasm population structure*"). For TMS13 target individuals, we listed all GG individuals as candidate parents. For TMS14 target individuals, we listed all GG and TMS13 individuals as candidate parents. For TMS15 target individuals, we listed all GG, TMS13, and TMS14 individuals as candidate parents.

Table 3.2 Results from AlphaAssign.
The table shows the results of our pedigree validation procedure. Rows highlighted in yellow represent useable data (either duos or trios). An individual's data is labeled "missing" when we either could not find GBS data for that individual or when we could not resolve their BIGRED results.

	GG	TMS13	TMS14	TMS15
Neither parent validated	43	197	361	765
One parent validated (useable as duos)	19	532	715	684
Both parents validated (useable as trios)	9	1524	1196	470
Missing data for one parent and the other parent was not	78	33	97	44

validated				
Missing data for one parent and the other parent was validated (useable as duos)	38	33	137	122
Missing data for both parents	54	0	0	4

Filtering the GBS allele depth data before calling genotypes

We have allelic depth (AD) data for each individual at each site. The AD data for individual *d* at site *v* is a record of the observed counts of each of the two alleles in individual *d* at site *v*: $X_d^{(v)} = \left(N_A^{(v,d)}, N_B^{(v,d)}\right)$, where $N_A^{(v,d)}$ and $N_B^{(v,d)}$ denote the observed counts of allele A and allele B, respectively, in individual *d* at site *v*. We removed sites with >70% missing data then calculated the proportion of missing data for each individual and removed individuals with >80% missing data. Here, we defined "missing" as observing zero reads for a given individual at a given site. The filter removed one individual IITA-TMS-IBA011610 from analysis. Exclusion of this individual causes offspring IITA-TMS-IBA062021 to have no listed father or mother. We included IITA-TMS-IBA062021 in the analysis when phasing and imputing. Removal of this duo is inconsequential since this duo provides an uninformative meiosis (see the section "*Filtering the SHAPEIT2-duoHMM output*" below for a discussion of informative meioses). We then removed sites with a mean depth greater than 120 to avoid spurious genotype calls within repeat regions, i.e., paralogs.

Generating input data files for SHAPEIT2

SHAPEIT2 takes called genotypes as input. To obtain a set of called genotypes for our sample, we first calculated genotype posterior probabilities for each individual at each site. Given observed data $X_d^{(v)}$ and fixed sequencing error rate e = 0.01, we computed the likelihood for genotype $G_d^{(v)} = g$. We calculated genotype likelihoods for a single individual at a single site, independent of all other individuals and sites in the sample, using the following equation:

$$P\left(X_d^{(v)} | G_d^{(v)} = g, e\right) = \binom{N_A^{(v,d)} + N_B^{(v,d)}}{N_B^{(v,d)}} (1 - p_B)^{N_A^{(v,d)}} (p_B)^{N_B^{(v,d)}} \quad (2)$$

$$p_B = \begin{cases} e, & \text{when } g = AA \\ 0.50, & \text{when } g = AB \\ 1 - e, & \text{when } g = BB \end{cases}.$$

We estimated posterior probabilities for the three genotypes using the likelihoods defined above and assuming a genotype prior (then normalizing by some constant). This genotype prior varied depending on whether individual *d* had zero validated parents (i.e., was a founder), had one validated parent, or had two validated parents. If individual *d* had zero validated parents, we calculated its genotype prior for site *v* using the estimated frequency of the reference allele at site *v* and assuming HWE. If individual *d* had one validated parent, we calculated its genotype prior for site *v* using the posterior probability distribution of its known parent, the genotype probability distribution of a 'dummy' parent, and the rules of Mendelian inheritance. We calculated the genotype probability distribution of the dummy parent at site *v* by

using the estimated frequency of the reference allele at site *v* and assuming HWE. If individual *d* had two validated parents, we used the posterior probability distributions of its known parents and Mendelian inheritance rules to calculate individual *d*'s prior. Notice that this scheme requires calculation of posterior genotype probabilities in a sequential manner, propagating information down the pedigree to subsequent generations.

We called genotypes from these estimated posterior genotype probabilities, calling a genotype for individual *d* at site *v* only if one of the three possible genotypes had a posterior probability greater than or equal to 0.99. To qualitatively examine how SHAPEIT2 performs at different levels of missing data, we generated seven datasets: datasets where we removed sites with more than 20%, 30%, 40%, 50%, 60%, and 70% missing data. We observed that when more markers are retained, SHAPEIT2-duoHMM detected a larger number of crossovers but crossover intervals were longer. Results from the 20% dataset were very noisy, so we selected the 30% dataset to analyze. Table 3 shows the number of sites after applying the 30% maximum-missing filter for each chromosome. Appendix Figure 3.1 shows the plots for each chromosome's duoHMM-inferred crossover intervals for the 20% and 30% maximum-missing datasets. When detecting recombination events using duoHMM, the algorithm returns a file with five columns: "CHILD", "PARENT", "START", "END", and "PROB_RECOMBINATION". The first two columns show the child and parent involved in the meiosis. The "START" and "END" columns define the regions (in bp) where a crossover may have occurred, and the final column lists the probability that a crossover event occurred in a given interval. Appendix Figure 3.1 shows those

crossover intervals with probabilities greater than or equal to $t = 0.9$ (refer to the section "Detecting recombination events using duoHMM" for a full description of duoHMM).

.

Table 3.3 Number of sites remaining after filtering
The table lists the number of sites in the dataset before and after application of the 30% maximum-missing filter for each chromosome and removing monomorphic and singleton sites.

Chromosome	Number of sites before filter	Number of sites after 30% maximum-missing filter	Number of sites after removing monomorphic and singleton sites
1	30575	17159	3739
2	24151	14818	2265
3	22156	14493	2458
4	18271	9980	2519
5	20750	14106	1681
6	20174	12907	1609
7	12226	7163	1114
8	17991	11549	1689
9	18267	12194	1341
10	14985	8206	1573
11	18816	11267	1777
12	15906	9921	1703
13	15851	9969	2018
14	20635	11957	2939
15	20048	13239	1705
16	15447	9832	1267
17	15390	8716	2206
18	15053	9063	1524

Implementation of SHAPEIT2 and duoHMM

We used SHAPEIT2 and duoHMM to detect SNP intervals flanking a crossover event (a recombination event can only be resolved down to the region between its two flanking heterozygous markers in the parent). We followed the

recommendations of O'Connell et al. [8] to phase and impute individuals in our pedigreed population. We ran SHAPEIT2, ignoring all explicit family information then applied duoHMM to combine the SHAPEIT2-inferred haplotypes with verified family information to correct switch errors (SEs). We carried out both steps internally within SHAPEIT2 by using the '—duohmm' flag.

SHAPEIT2 combines features of SHAPEIT1 and Impute2. Specifically, SHAPEIT2 uses the SHAPEIT1 HMM to represent the space of haplotypes consistent with a given individual's genotypes across a chromosome with the difference being that in SHAPEIT2, the transition probabilities of the HMM are estimated by applying the Impute2 'surrogate family' phasing approach in local windows of size W. Under this scheme, K informative haplotypes are chosen to update the transition probabilities of the HMM in each window. We describe the SHAPEIT1 HMM below.

Suppose we have a sample of n diploid individuals. Let $G = (G_1, \ldots, G_n)$ denote the (observed) genotypes of the n individuals at L SNPs, i.e. $G_1 = (g_1, \ldots, g_L)$. The total possible number of distinct haplotype pairs consistent with genotype G_i is equal to $2^{(z-1)}$, where z denotes the number of heterozygous SNPs present in G_i. Let S_i represents the space of possible haplotype pairs consistent with G_i and $S = (S_1, \ldots, S_n)$ denote the total haplotype space for the n individuals. Let $\rho = (\rho_1, \ldots, \rho_{L-1})$ denote the vector of recombination rates between each pair of consecutive SNPs as described by Stephens and Scheet [11]. Let $H = (H_1, \ldots, H_n)$ denote the true (unobserved) haplotype pairs corresponding to $G = (G_1, \ldots, G_n)$. SHAPEIT regards H as unobserved random quantities with sampling space in S and aims to estimate the posterior distribution of H given G and ρ.

Because $\Pr(H|G, \rho)$ cannot be calculated exactly, SHAPEIT1 uses Gibbs sampling, a type of Markov chain-Monte Carlo (MCMC) algorithm, to approximate it by obtaining an approximate sample from the posterior distribution. The algorithm starts with an initial guess for H and a random order of treatment for the n individuals, ordering v. To iterate from $H^{(t)}$ to $H^{(t+1)}$, SHAPEIT1 updates the haplotype pair of each individual i in turn (in the order given by v) by sampling from the conditional distribution $\Pr(H_i|G, H_{-i}^{(t)}, \rho)$, where $H_{-i}^{(t)}$ is the set of current guesses for the haplotypes of all individuals except i. The conditional distribution $\Pr(H_i|G, H_{-i}^{(t)}, \rho)$ depends on assumptions about the genetic and demographic history underlying the data, i.e. a "prior" for the population haplotype frequencies. SHAPEIT1 infers haplotypes under the genetic model of coalescence with recombination developed by Stephens and Donnelly [12], which employs a "coalescent with recombination" prior to reflect the fact that each sampled haplotype will be similar to another haplotype or be a mosaic of other haplotypes in the pool of $2n$-2 haplotypes, altered by mutations and recombination, respectively. An iteration of the Gibbs sampling procedure completes when all n individuals have been updated. At the end of each iteration, SHAPEIT1 accepts or rejects new values for ρ and v according to the Metropolis-Hastings acceptance probability. Repeating this process enough times results in an approximate sample from $\Pr(H|G, \rho)$.

SHAPEIT1 computes $\Pr(H_i|G, H_{-i}^{(t)}, \rho)$ via implementation of an HMM, where the $2n$-2 haplotypes represent the hidden states of the HMM, ρ encodes the transition probabilities, and a constant mutation parameter encodes the emission

probabilities. To compute the probability of observing h, one must sum up the probabilities of observing h over all $(2n\text{-}2)^s$ possible sequences of s different states. This is done efficiently by implementation of the forward algorithm. SHAPEIT1 and PHASE2 differ in how they represent the space of possible haplotype pairs S_i. PHASE2 computes $\Pr(H_i|G, H_{-i}^{(t)}, \rho)$ from a complete haplotype list, whereas SHAPEIT1 computes the distribution from an incomplete binary tree, permitting the use of the PHASE2 model on larger datasets.

SHAPEIT2 takes as input a set of genotypes and a genetic map. SHAPEIT2 outputs either a single set of estimated (most-likely) haplotypes or a haplotype graph that encapsulates the uncertainty about the underlying haplotypes. We chose the latter output. SHAPEIT2 has multi-threading capabilities, but we chose not to use this feature in order to maximize the number of individuals that SHAPEIT2 conditions on during Gibbs sampling. When running SHAPEIT2 using four threads on a dataset of 100 individuals, the algorithm will phase four individuals simultaneously, conditional upon the 100-4 other individuals in the dataset. We ran SHAPEIT2 with 14 burn-in iterations, 16 pruning iterations, and 40 main iterations. We increased the number of conditioning states to 200 states per SNP. The developers found it slightly advantageous to use a window size larger than 2 Mb when large amounts of identical by descent (IBD) sharing are present. We used a window size of 5 Mb. We provided SHAPEIT2 a genetic map that specifies the recombination rate between SNPs. We generated this genetic map by interpolating genetic distances of GBS markers using ICGMC's composite genetic map. We used

the default value of 15,000 for the effective population size, a parameter that scales the recombination rates that SHAPEIT2 uses to model patterns of LD.

The goal of duoHMM is to detect the genuine recombination events and correct SEs in SHAPEIT2-inferred haplotypes. Let S_v denote the pattern of gene flow at locus v, where $S_v = (j,k)$ denotes the scenario where parental haplotype j and child haplotype k are IBD. Here $j,k \in \{1,2\}$, so there are four possible patterns of gene flow between a parent and child at a site v:

1. the allele on parental haplotype 1 and the allele on child haplotype 1 are IBD, denoted $S_v = (1,1)$ or A,
2. the allele on parental haplotype 2 and allele on child haplotype 1 are IBD, denoted $S_v = (2,1)$ or B,
3. the allele on parental haplotype 1 and allele on child haplotype 2 are IBD, denoted $S_v = (1,2)$ or C, and
4. the allele on parental haplotype 2 and allele on child haplotype 2 are IBD, denoted $S_v = (2,2)$ or D.

The true pattern of gene flow at each site is unobserved, and duoHMM infers the true inheritance states from the (imperfect) observed parental and child haplotypes, i.e., SHAPEIT2-inferred haplotypes. Because the rate of recombination in any given meiosis is low, the HMM is parameterized such that the pattern of gene flow remains constant over long stretches of a chromosome. Figure 3.1 enumerates and explains the possible transitions from site v to site v+1 for the case where $S_v = (1,1)$. We refer

the reader to the duoHMM paper for full specification of the HMM (enumeration of all possible transition types, definition of transition rates, estimation of parameters) [8]. After estimating parameters of the HMM using the Forward Backward algorithm, duoHMM finds the most likely state sequence using the Viterbi algorithm. When duoHMM infers a SE in the Viterbi sequence in either the parent or child, duoHMM corrects the haplotypes by switching the phase of all loci proceeding the SE. The algorithm applies these corrections sequentially down through each pedigree.

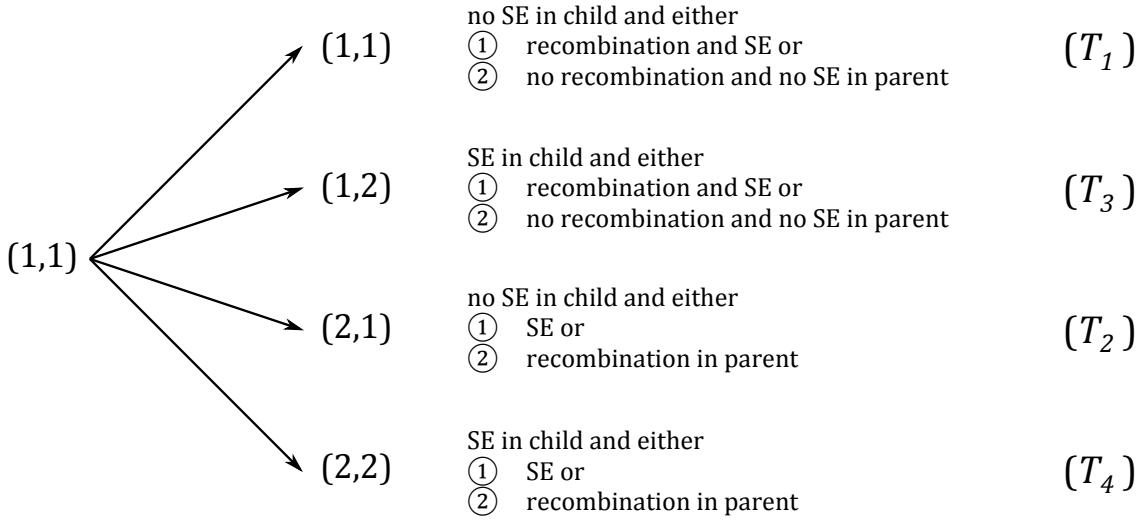


Figure 3.1 Possible inheritance state transitions from site *v* to site *v*+1 for the case where S_v = (1,1) for duoHMM.
When duoHMM observes a T3 or T4 transition in the Viterbi sequence, it infers a SE in the child haplotypes. When duoHMM observes a T2 or T4 transition, it infers either a SE or a recombination event in the parental haplotypes, but determining which of the two events actually occurred is difficult. The algorithm makes a decision by looking at all the offspring of that parent. When one of the T2 or T4 transitions is present in the same location for the majority of offspring, the transition is most likely a SE on the parental haplotypes (minimum-recombinant solution).

Detecting recombination events using duoHMM

Once SHAPEIT-inferred haplotypes had been corrected with duoHMM, we reran duoHMM to infer recombination events. The HMM infers recombination events by calculating the probability of a recombination event between markers. We refer the reader to the duoHMM paper for an explaination of how this probability is calculated [8]. To detect crossovers, we sampled a haplotype pair for each individual from SHAPEIT2's diploid graph then calculated the probability of a recombination event between pairs of markers. We repeated this process a total of 10 times then averaged the inter-SNP recombination probabilities across the 10 iterations. We included a crossover interval in subsequent analyses if the interval had a probability greater than or equal to t. We set $t = 0.5$.

Filtering the SHAPEIT2-duoHMM output

The power to detect recombination events is dependent on the structure of the pedigree. In a nuclear family with >2 offspring, most crossover events should be detectable, and we classify these pedigrees as *informative* towards recombination. We analyzed data from only those pedigrees having "informative" meioses, which we defined as a nuclear family consisting of >2 offspring or a pedigree consisting of three generations. We refer to the parents of these pedigrees as "informative parents" and the meioses in these pedigrees as "informative meioses". Of the total 8,678 meioses in the data set, 7,165 were informative (3,679 female meioses; 3,486 male meioses).

Building the genetic maps

To build a genetic map, we first calculated the number of crossover events between each inter-SNP interval. If a crossover event spanned multiple SNP intervals, we assigned a fraction of the crossover event to each of the spanned intervals, calculated as one divided by the length of the inter-SNP interval in base pairs. We then calculated the genetic length of each SNP interval on chromosome *y* by dividing the number of crossovers in each interval by *n*, where *n* = (the genetic length of chromosome *y* in the ICGMC map)/(the total number of crossovers we detected on chromosome *y*). We did this so that our genetic map for each chromosome ends at the same genetic position as ICGMC's map.

Examining evidence of sexual dimorphism

We next examined the distribution of crossover events along each chromosome for female and male meioses, separately. We divided each chromosome into windows of 1-Mb and determined the number of male meiotic crossovers and female meiotic crossovers in each window. To examine if crossover counts in each window varied between the sexes, we performed a chi-square test of equal counts in each window. To calculate the expected number of male crossovers in a given window, we calculated the proportion of total meioses analyzed that were male (i.e., 3,486/(3,6789 + 3,486)) then multiplied this value by the total number of crossovers in the window. We calculated the expected number of female crossovers in a given window in the same way. We did not test for statistical significance in the

last window of any chromosome since the last window is shorter than 1-Mb (no chromosome is perfectly divisible by 1-Mb). We tested each window at a Bonferroni-corrected significance level of α/m, where $\alpha = 0.05$ and $m = 506$ (i.e., the total number of windows tested). We also performed this test genome-wide at a significance level of 0.05.

Examining if crossover placements are random and independent events

If crossover placements are random and independent events, the distribution of the number of crossovers observed on a given chromosome in a given parent-offspring pair is expected to follow a Poisson distribution. We used the deviance goodness of fit test to test if crossover placements are random and independent events. For each chromosome, we performed a Poisson regression where we modeled the number of crossovers observed in a given parent-offspring pair Y as a function of the covariates "parent" and "sex". The "parent" covariate specifies the parent involved in the parent-offspring pair, and the "sex" covariate species whether the parent was a female or male (i.e., where the crossovers observed in a male or female meiosis). We used the residual deviance to perform a chi-square goodness of fit test for the overall model. The residual deviance is the difference between the deviance of the current model and the maximum deviance of the ideal model where the predicted values are identical to the observed. If the residual difference is small enough, the goodness of fit test will not be significant, indicating that the Poisson model fits the data. We performed these test at a Bonferroni-corrected significance level of α/m, where $\alpha = 0.05$ and $m = 18$ (i.e., the total number of chromosomes tested).

RESULTS

Using SHAPEIT2 and duoHMM, we detected a total of 65,771 and 65,287 crossover-containing intervals from female and male meioses, respectively, across the 18 chromosomes. Using these crossover intervals, we constructed a sex-averaged genetic map, which we compared to an existing map constructed by ICGMC, and sex-specific genetic maps. Our sex-averaged map has a median resolution of 420,366 bp. The female and male genetic maps have median resolutions of 397,433 bp and 433,827 bp, respectively.

To compare our map to ICGMC's, we plotted the genetic position (cM) of our markers and ICGMC's markers as a function of physical position (Mb). Figure 3.2 shows the results for chromosomes 1 and 4. We show the plots for each chromosome in Appendix Figure 3.2. At the qualitative level, the distribution of crossovers observed in our map is in good agreement with the ICGMC map.

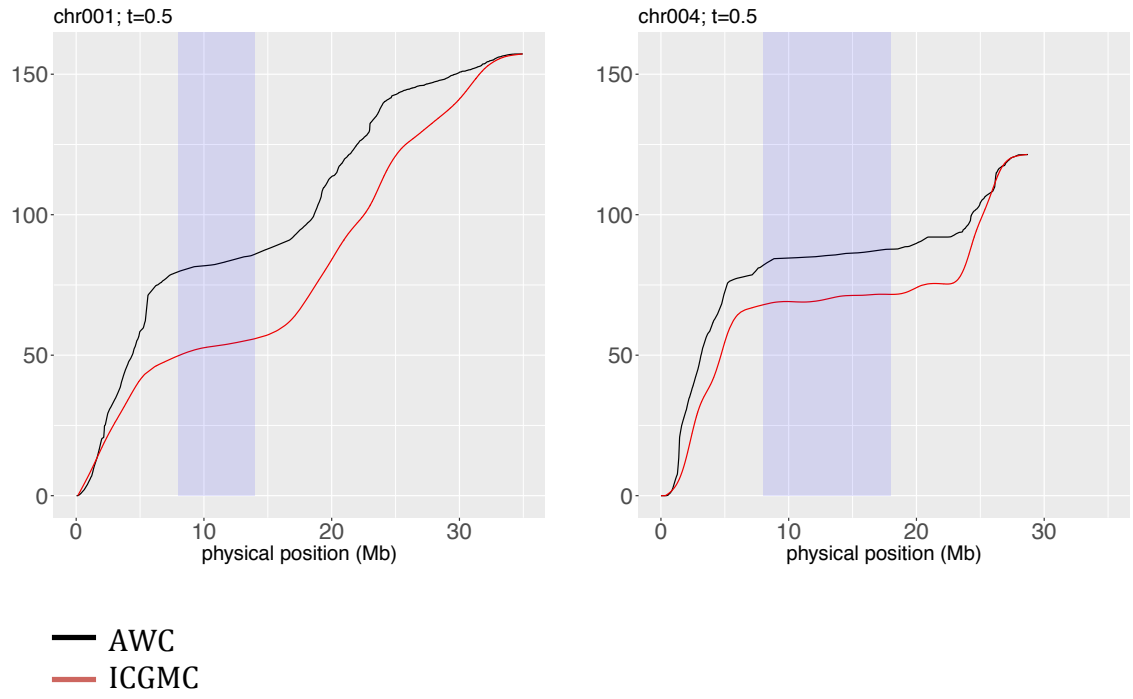


Figure 3.2 Comparison of our genetic map (AWC) with ICGMC's.
We plotted the genetic position of our GBS markers (black) and ICGMC's markers (red) as a function of physical position (Mb) for chromosomes 1 and 4. The left plot shows the comparison for chromosome 1 and the right for chromosome 4. Centromeric regions of chromosomes are shaded in blue.

We plotted the number of crossover events in 1-Mb windows along each chromosome. Figure 3.3 shows the plot for chromosome 1. Appendix Figure 3.3 shows these plots for all 18 chromosomes. We found that crossovers are suppressed around centromeric regions of chromosomes. The correlation between the number of crossovers on each chromosome and the physical size of each chromosome was 0.40. We next examined the distribution of crossovers along each chromosome for female and male meioses, separately. We again divided each chromosome into windows of 1-Mb and plotted the number of crossovers detected in female meioses and male meioses in each 1-Mb window (Fig 3.3; red and blue). The spatial distribution of crossovers along the chromosomes does not vary between male and

female meiosis. To examine if crossover frequency in each window varied between the sexes, we performed a chi-square test of equal counts in each window. We did not test for statistical significance in the last window of any chromosome since the last window is shorter than 1-Mb. Of the 506 intervals tested, 45 (8.89%) passed the significance threshold. In these 45 intervals, female crossover count was significantly higher than that observed in males. Statistically significant intervals did not consistently appear in any specific region of chromosomes (Appendix Fig 3.3).

We tested if there is sexual-dimorphism in crossover number at the genome-wide level and found that the number of crossovers observed in male and female meioses significantly differed (p-value < 2.2×10^{-16}).

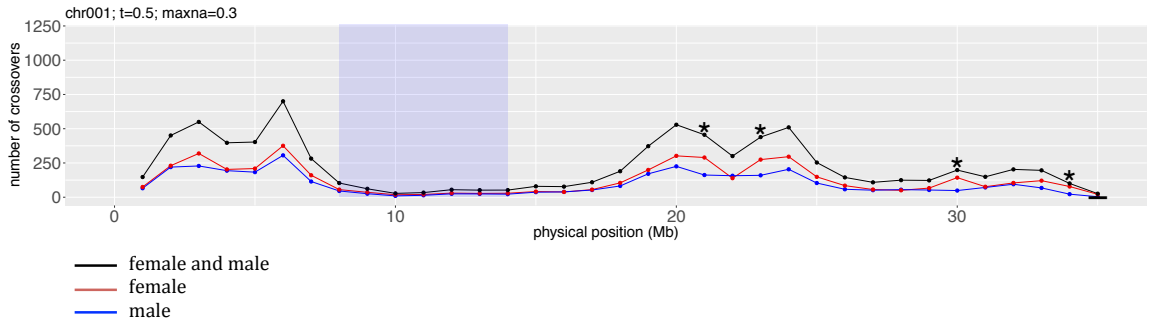


Figure 3.3 Distribution of crossover events across chromosome 1 for all meioses, female meioses, and male meioses.
We divided each chromosome into 1-Mb windows and plotted the number of crossovers falling within each interval for all (black), female (red), and male (blue) meioses. Asterisks show intervals with significantly different crossover frequency between male and female meioses. Dashes represent cases where we could not perform the chi-square test because the expected frequency count for one or more classes was less than five. We did not test for statistical significance in the last window of any chromosome since the last window is shorter than 1-Mb (no chromosome is perfectly divisible by 1-Mb). These intervals are annotated with a dash. The centromere of chromosomes is shown in blue. We tested each interval at a significance level of α/n, where $\alpha = 0.05$ and $n = 506$.

We used the deviance goodness of fit test to test if crossover placements are random and independent events. The goodness of fit test was significant for all chromosomes except chromosomes 10, 17, and 18, indicating that the Poisson model does not fit the data observed on chromosomes 1-9 and 11-16 well.

DISCUSSION

We used IITA's multi-generational pedigree, consisting of 7,165 informative meioses (3,679 female; 3,486 male), to characterize recombination in cassava. Using SHAPEIT2 and duoHMM, we detected a total of 65,771 and 65,287 crossover-containing intervals from female and male meioses, respectively, across the 18 chromosomes. Using these crossover intervals, we constructed a genetic map and compared it to an existing map constructed by ICGMC. To study recombination differences between the sexes, we compared crossover number and spatial distribution along the 18 chromosomes between the sexes.

We observed similar spatial distributions of crossover events between the ICGMC map and our map, although it should be noted that we used a version of the ICGMC map as input when running SHAPEIT2 and duoHMM. Although not ideal, the ICGMC map only served as a prior for the SHAPEIT2 HMM, and recombination rates between SNPs were updated at the end of each iteration of Gibbs sampling. The Stephens and Donnelly model are also not sensitive to initial values. Additionally, two ICGMC parents were also parents in our pedigree. In any case, a less confounding comparison can be made with the results published by Ramu et al., where they analyzed recombination in 241 diverse accessions [13]. We recovered

similar patterns of crossover distribution in light of using different germplasm, suggesting that recombination is stable among different lines of cassava.

Our map does a slightly better job at capturing the linkage disequilibrium in regions that we know have non-recombining introgressed segments relative to ICGMC's (Fig 3.2). In the 1930's, breeders crossed cassava with its wild relative *Manihot glaziovii* in an effort to introduce cassava mosaic disease resistance into cassava. Marnin et. al found long segments of *M. glaziovii* haplotypes in modern cassava germplasm on chromosomes 1 and 4 [14]. The largest introgressions were detected on chromosome 1, spanning from 25 Mb to the end of the chromosome, and on chromosome 4 from 5 Mb to 25 Mb, both of which our map captures.

Differences between our map and ICGMC's could result from a number of reasons. The data used in our analysis was generated using a substantially different variant discovery pipeline than that used by ICGMC [15],[7]. We found only 97 SNPs (summed across all 18 chromosomes) in common between our map and ICGMC's. The ICGMC map was generated using 10 nuclear families, each family consisting of 117 to 256 offspring. Our map was generated using two multi-generational families (a family was defined as all individuals reachable in the pedigree graph through either ancestors or descendants), one family consisting of 4175 family members and the other consisting of 22 family members. There is also the question of what value of *t* to use, as this dictates the number of crossovers available for map building. Using a higher *t* value results in more confident crossover intervals but also a lower number of crossovers.

The regions of suppressed recombination on chromosome 1 and 4 observed in our genetic map coincide with the location of these introgressed segments, supporting the hypothesis that there is some mechanism keeping beneficial alleles in cis. It would be interesting to plot the distribution of crossovers separately for *M. esculenta* with zero, one, and two copies of these introgressions. One would think that individuals carrying one copy of the introgression would have suppressed recombination in that region relative to individuals with zero and two copies. Since a minimum number of crossovers must occur on a given tetrad for proper chromosomal segregation, it would be interesting to see if there are more recombination events upstream of the introgression in individuals with one copy of the introgression relative to individuals with zero and two copies.

In this study, we used the multi-generational pedigree and GBS data from IITA to sudy recombination in cassava. We characterized the recombination landscape across the 18 chromosomes of cassava and found that crossover rates vary greatly along the chromosomes and that all chromosomes except chromosomes 10, 17, and 18 displayed crossover interference. We constructed a genetic map using duoHMM-corrected, SHAPEIT2-inferred crossover intervals and compared it to ICGMC's composite map. We also examined female and male meioses, separately and found evidence that female meioses undergo more recombination than male meioses. The spatial pattern of crossovers along the chromosomes, however, does not vary between male and female meiosis at the qualitative level.

REFERENCES

[1] D. P. Moore and T. L. Orr-Weaver, “8 Chromosome Segregation during Meiosis: Building an Unambivalent Bivalent,” *Curr. Top. Dev. Biol.*, 1997.

[2] C. Larchevêque, J. B. Fernandes, M. Séguéla-Arnaud, A. H. Lloyd, and R. Mercier, “Unleashing meiotic crossovers in hybrid plants,” *Proc. Natl. Acad. Sci.*, 2017.

[3] K. R. Ritz, M. A. F. Noor, and N. D. Singh, “Variation in Recombination Rate: Adaptive or Not?,” *Trends in Genetics*. 2017.

[4] L. H. Hartwell, L. Hood, M. L. Goldberg, A. E. Reynolds, and L. M. Silver, *Genetics: From Genes to Genomes*, 4th ed. 2011.

[5] T. Lenormand and J. Dutheil, “Recombination difference between sexes: A role for haploid selection,” in *PLoS Biology*, 2005.

[6] F. Awoleye, M. Vanduren, J. Dolezel, and F. J. Novak, “Nuclear-DNA Content and in-Vitro Induced Somatic Polyploidization Cassava (Manihot-Esculenta Crantz) Breeding,” *Euphytica*, vol. 76, no. 3, pp. 195–202, 1994.

[7] “ High-Resolution Linkage Map and Chromosome-Scale Genome Assembly for Cassava (Manihot esculenta Crantz) from 10 Populations ,” *G3: Genes|Genomes|Genetics*, 2014.

[8] J. O’Connell, D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, M. Cocca, M. Traglia, J. Huang, J. E. Huffman, I. Rudan, R. McQuillan, R. M. Fraser, H. Campbell, O. Polasek, G. Asiki, K. Ekoru, C. Hayward, A. F. Wright, V. Vitart, P. Navarro, J. F. Zagury, J. F. Wilson, D. Toniolo, P. Gasparini, N. Soranzo, M. S. Sandhu, and J. Marchini, “A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness,” *PLoS Genet.*, 2014.

[9] A. W. Chan, A. L. Williams, and J. L. Jannink, “A statistical framework for detecting mislabeled and contaminated samples using shallow-depth sequence data,” *BMC Bioinformatics*, 2018.

[10] A. Whalen, G. Gorjanc, and J. M. Hickey, “Parentage assignment with genotyping-by-sequencing data,” *Journal of Animal Breeding and Genetics*, 2018.

[11] M. Stephens and P. Scheet, “Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation,” *Am. J. Hum. Genet.*, 2005.

[12] M. Stephens and P. Donnelly, “A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data,” *Am. J. Hum. Genet.*, 2003.

[13] P. Ramu, W. Esuma, R. Kawuki, I. Y. Rabbi, C. Egesi, J. V. Bredeson, R. S. Bart, J. Verma, E. S. Buckler, and F. Lu, “Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation,” *Nat. Genet.*, 2017.

[14] M. Wolfe, G. J. Bauchet, A. W. Chan, R. Lozano, P. Ramu, C. Egesi, R. Kawuki, P. Kulakow, I. Rabbi, and J.-L. Jannink, “Introgressed Manihot glaziovii Alleles in Modern Cassava Germplasm Benefit Important Traits and Are Under Balancing Selection,” 2019. *Submitted*

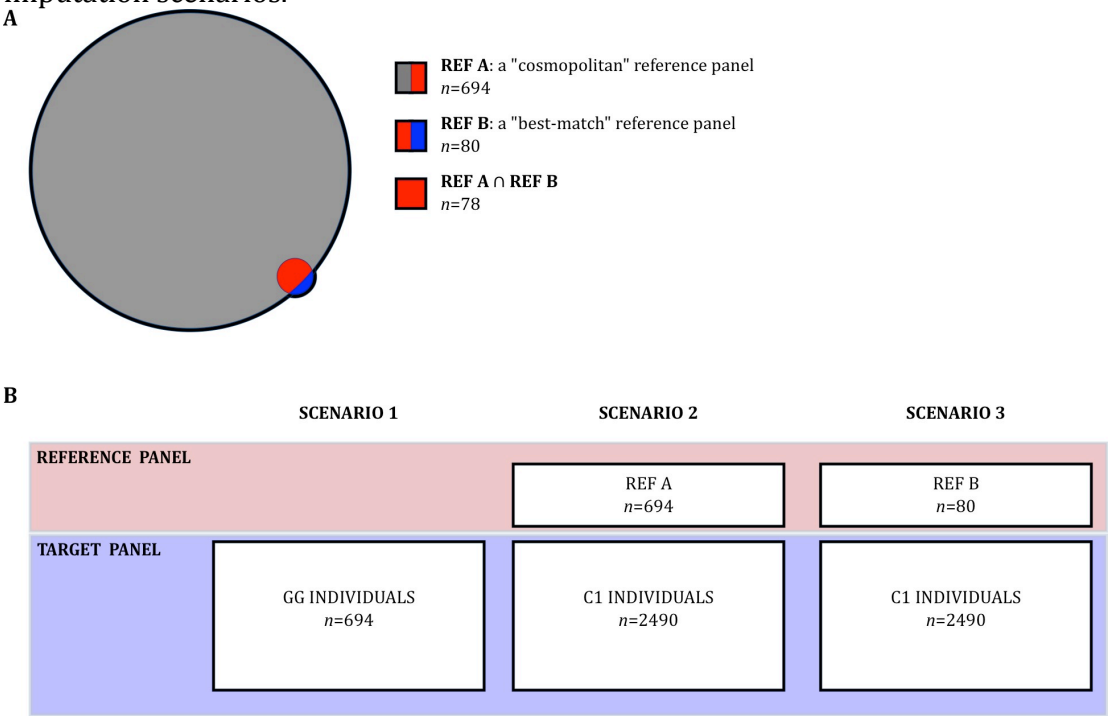
[15] A. W. Chan, M. T. Hamblin, and J.-L. Jannink, “Evaluating Imputation Algorithms for Low-Depth Genotyping-By-Sequencing (GBS) Data.,” *PLoS One*, vol. 11, no. 8, p. e0160733, 2016.

[16] J. V Bredeson, J. B. Lyons, S. E. Prochnik, G. A. Wu, C. M. Ha, E. Edsinger-Gonzales, J. Grimwood, J. Schmutz, I. Y. Rabbi, C. Egesi, P. Nauluvula, V. Lebot, J.

Ndunguru, G. Mkamilo, R. S. Bart, T. L. Setter, R. M. Gleadow, P. Kulakow, M. E. Ferguson, S. Rounsley, and D. S. Rokhsar, “Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity,” *Nat Biotech*, vol. advance on, Apr. 2016.

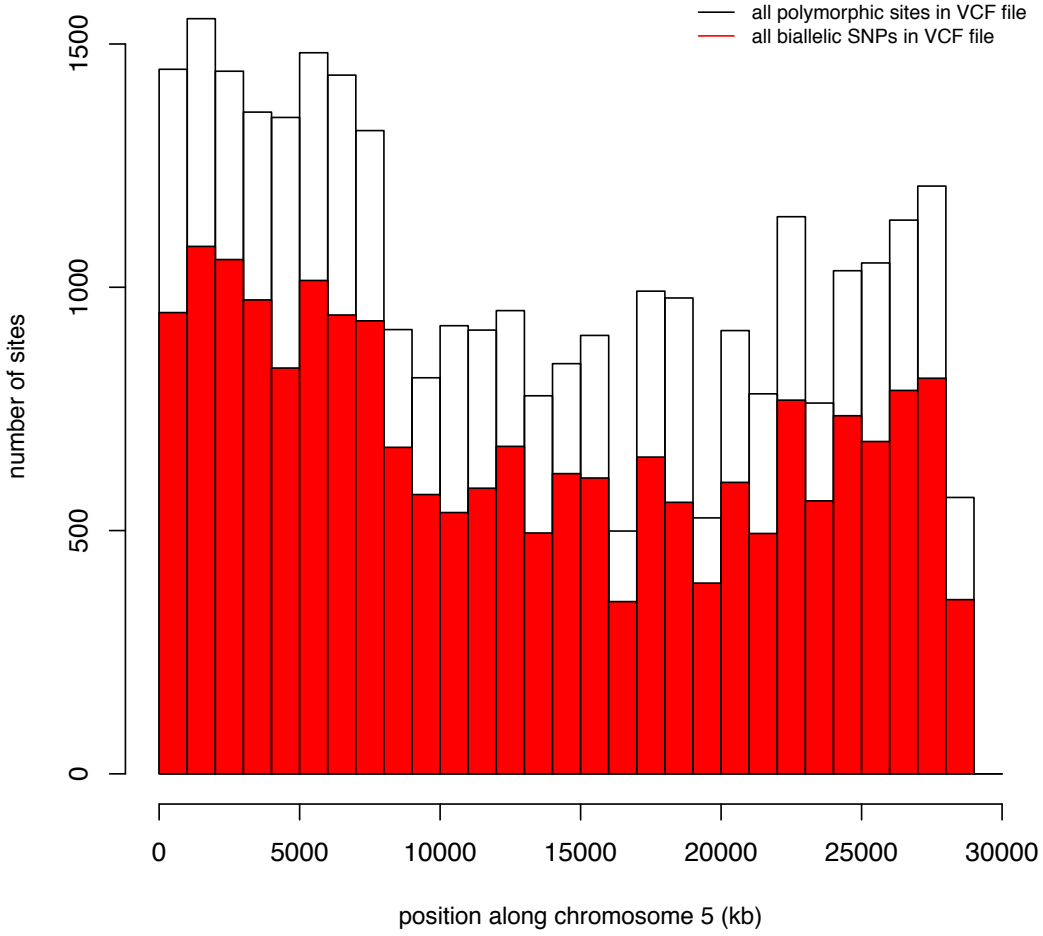
APPENDIX

Appendix Figure 1.1 Description of reference panel A and B and the three imputation scenarios.



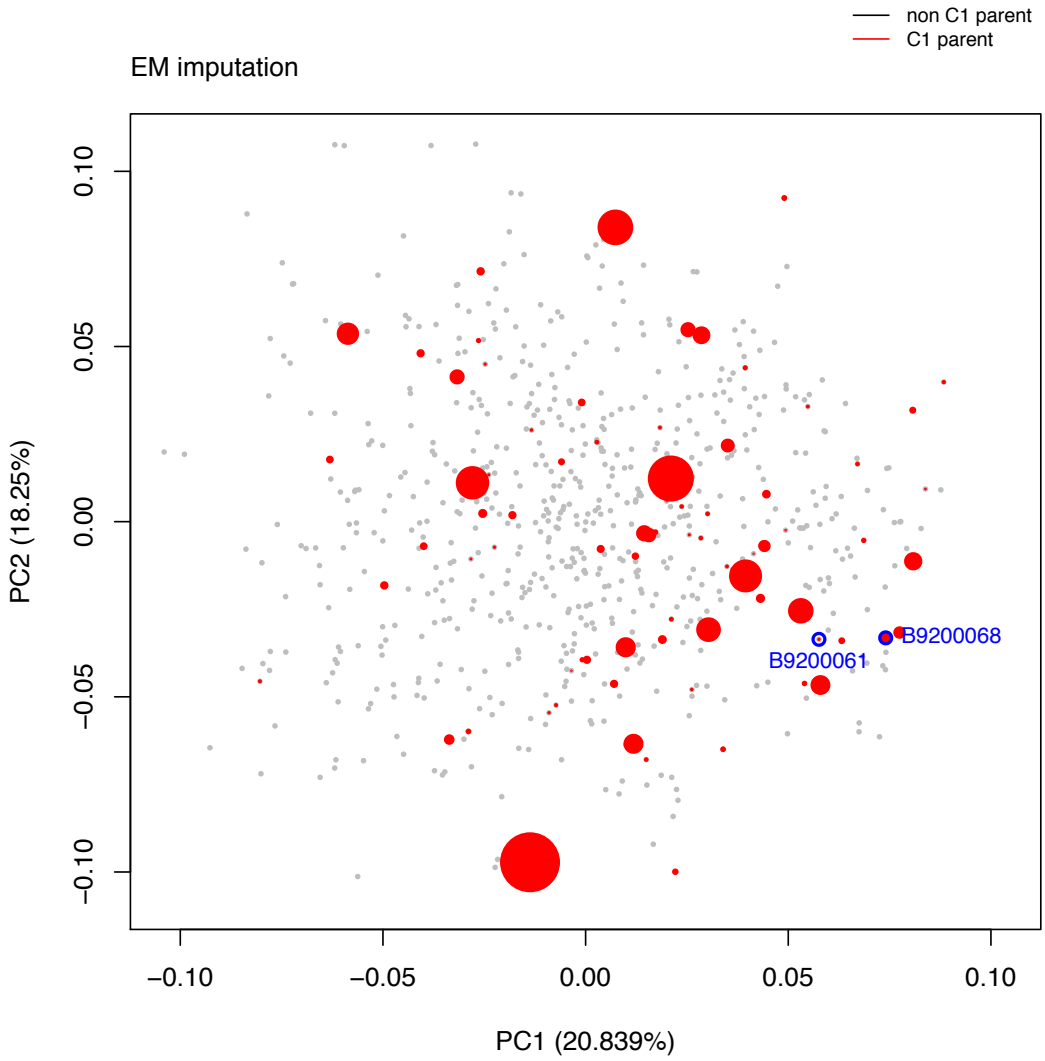
(A) The Venn diagram shows the composition of reference panel A and B. (B) We evaluated Beagle and glmnet under three imputation scenarios: imputation guided by no reference panel (left), a reference panel with large genetic diversity (reference panel A; middle), and 3) a reference panel that closely matches the ancestry of the study sample (reference panel B; right).

Appendix Figure 1.2 Distribution of variants across chromosome 5.



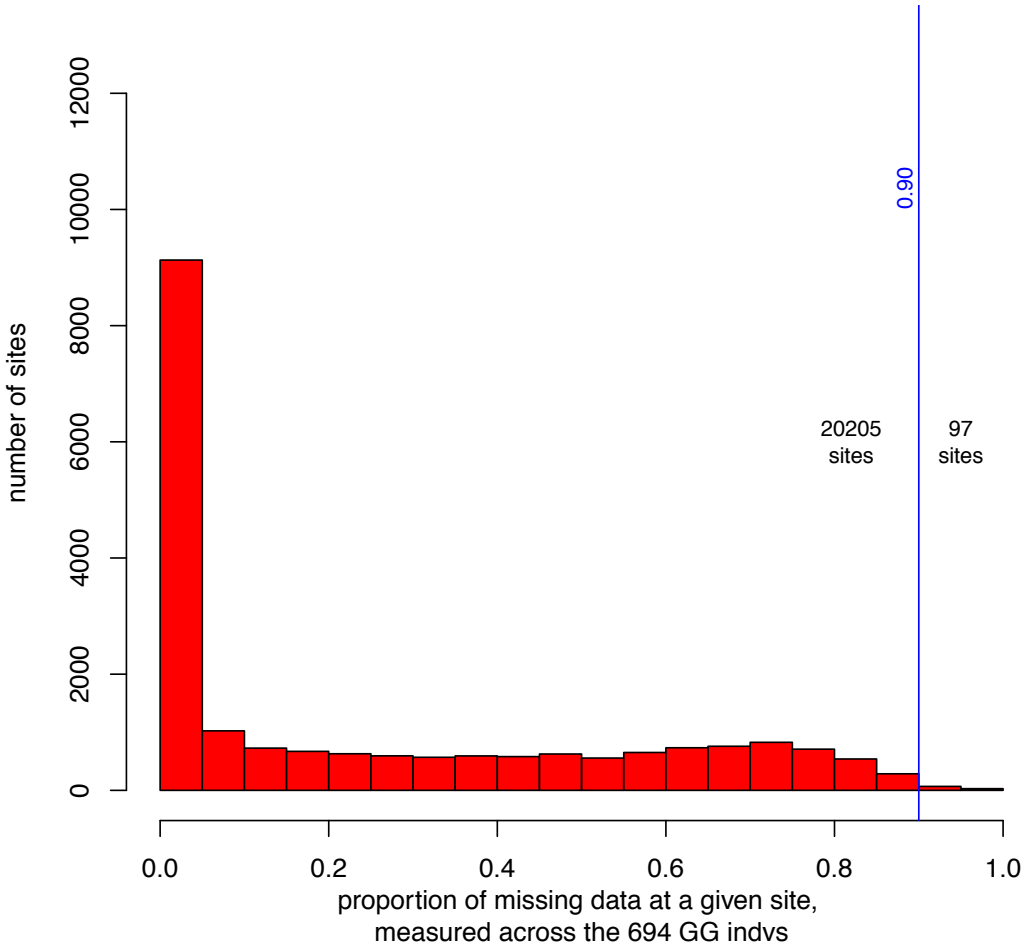
The white and red histogram displays the distribution of all variant sites (30018) and biallelic SNPs (20302) along the length of chromosome 5, respectively.

Appendix Figure 1.3 No evidence of population structure among the 696 reference panel individuals.



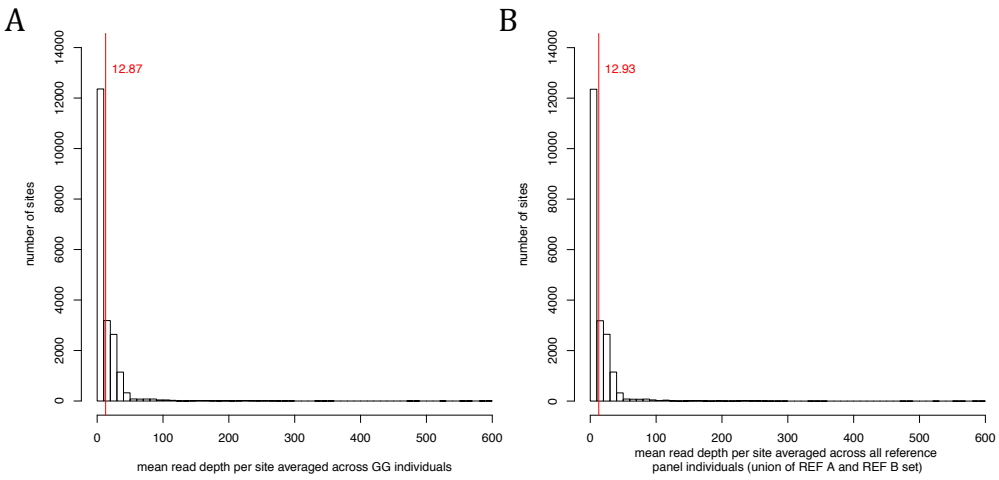
No records of genetic relatedness among the 696 reference panel individuals exist. We, therefore, performed a PCA to explore whether there is any evidence of population structure among reference panel individuals. Reference panel individuals contributing zero offspring to the C1 population appear as grey dots. Reference panel individuals contributing >0 offspring to the C1 population appear as red dots with diameters scaled proportionally to the number of offspring contributed by the individual.

Appendix Figure 1.4 Distribution of the proportion of missing data per biallelic SNP.



The proportion of missing data at a given site is measured across the 694 GG individuals. The term "missing" denotes zero reads observed at a given site for a given individual. We removed sites with >90% missing data, leaving a total of 20205 sites for cross-validation experiment 1. We used this same set of sites for scenarios 2 and 3 for reasons given in the main text.

Appendix Figure 1.5 Distribution of the mean read depth per site.



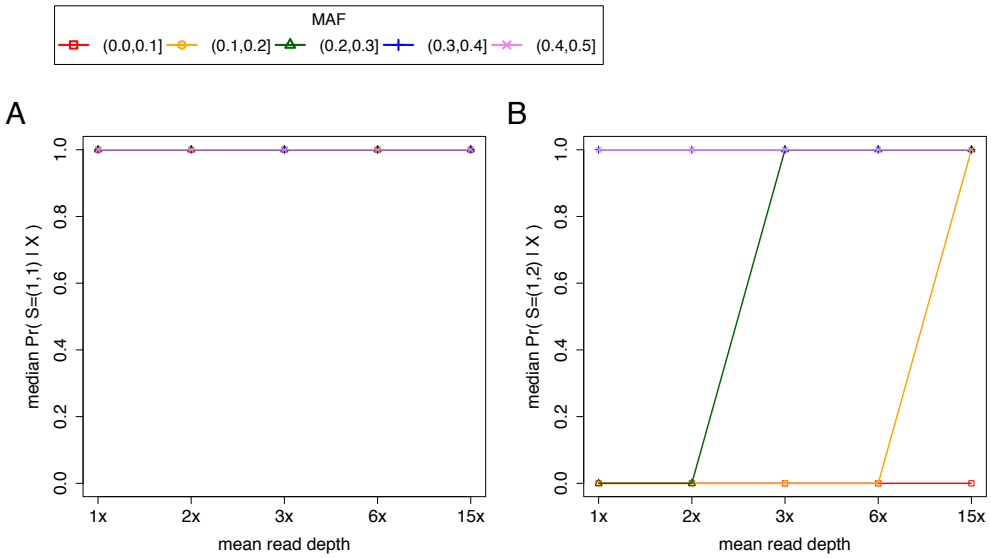
(A) The histogram shows the distribution of the mean read depth per site averaged across all 694 GG individuals. (B) The histogram shows the distribution of the mean read depth per site averaged across all 696 reference panel individuals. The red vertical line marks the mean of the distribution.

Appendix Figure 2.1 A simplified representation of a VCF data file containing allele depth (AD) data for the $k = 3$ putative replicates of I011206.

	$d = 1$	$d = 2$	$d = 3$
SNP 1	7,0	5,0	3,1
SNP 2	3,4	3,0	4,0
⋮	⋮	⋮	⋮
SNP M	4,0	5,1	7,0

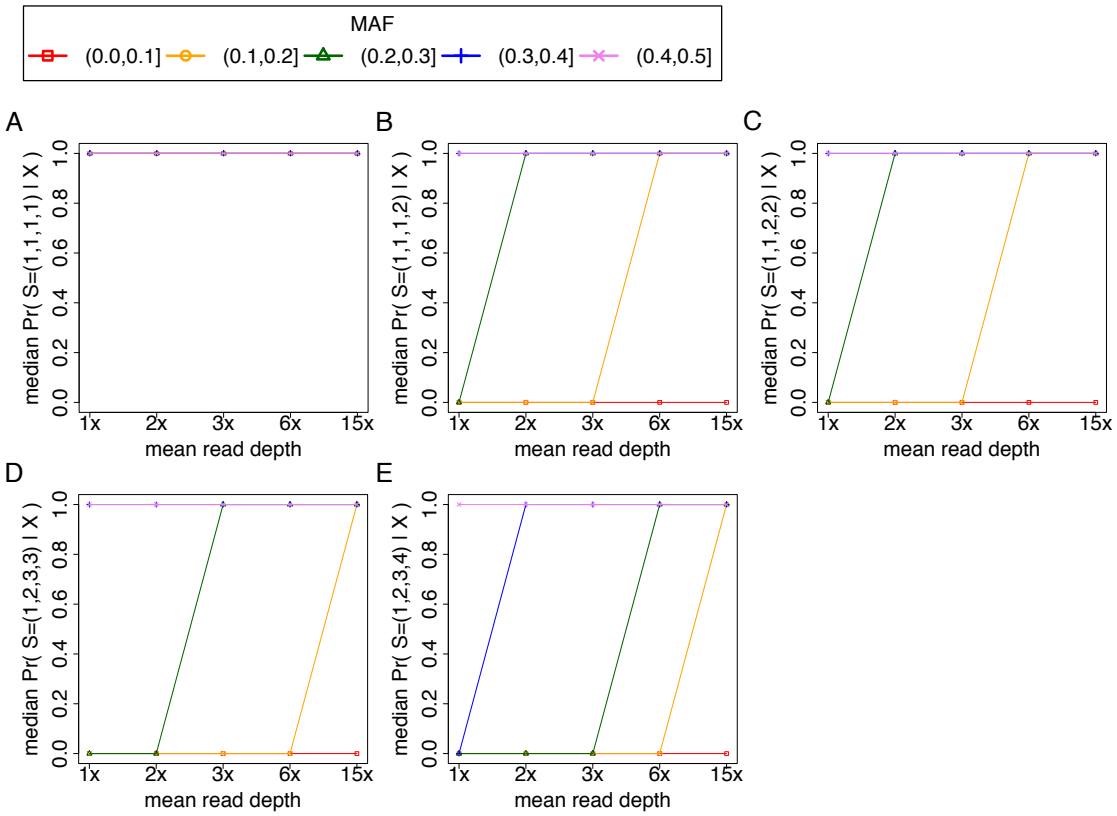
In this example, we have shallow sequenced individual I011206 three different times using some HTS method. We obtained this output: a matrix whose rows represent polymorphic sites and whose columns represent the replicate sequence runs. The putative replicates are indexed with the variable *d*. Each element of the matrix consists of two (comma-separated) integers, representing the observed counts for allele A and B. We wish to determine whether the DNA samples from these three sequence runs originate from one individual.

Appendix Figure 2.2 BIGRED's accuracy as a function of the mean read depth of samples and the MAF of analyzed sites for $k = 2$.



(A and B) Each plot shows estimates of the median posterior probability of the true source vector (y-axis) as a function of mean read depth of samples (x-axis) and MAF of sites (legend). Each data point presents the median posterior probability of S = (1,1) and S = (1,2) across 15 and 100 runs of the algorithm, respectively.

Appendix Figure 2.3 BIGRED's accuracy as a function of the mean read depth of samples and the MAF of analyzed sites for $k = 4$.

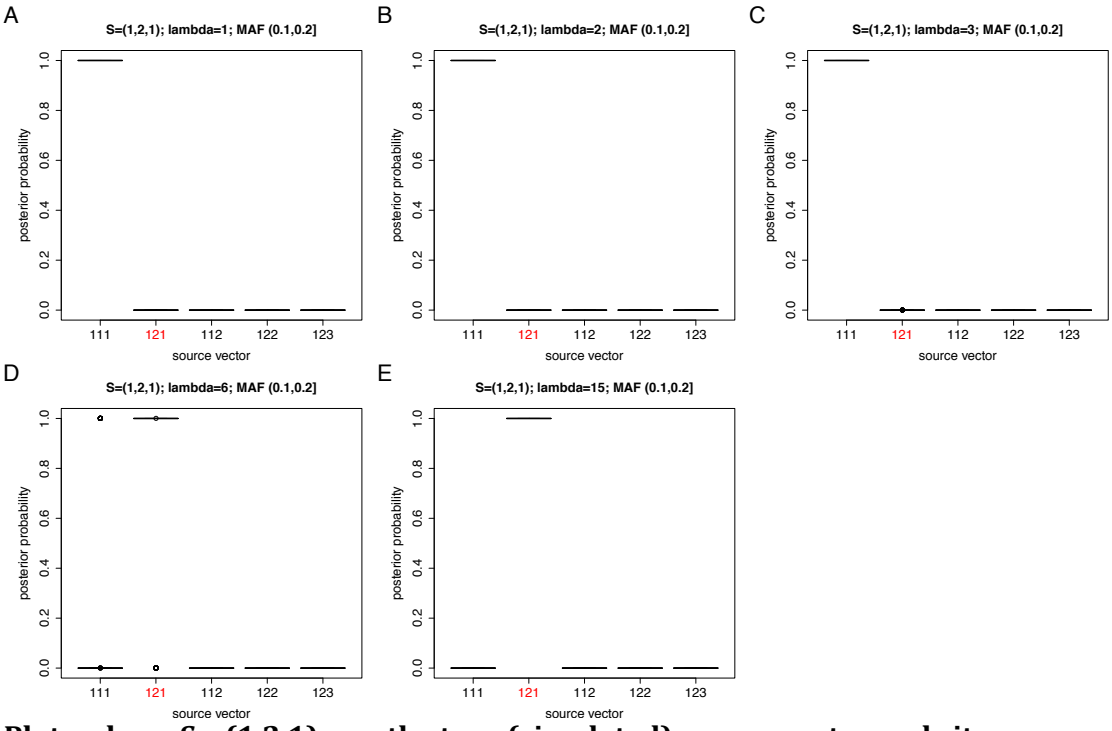


(A, B, C, D, and E) Each plot shows estimates of the median posterior probability of the true source vector (y-axis) as a function of mean read depth of samples (x-axis) and MAF of sites (legend). Each data point presents the median posterior probability of $S = (1,1,1,1)$, $S = (1,1,1,2)$, $S = (1,1,2,2)$, $S = (1,2,3,3)$, and $S = (1,2,3,4)$ across 15, 100, 100, 100, and 100 runs of the algorithm, respectively.

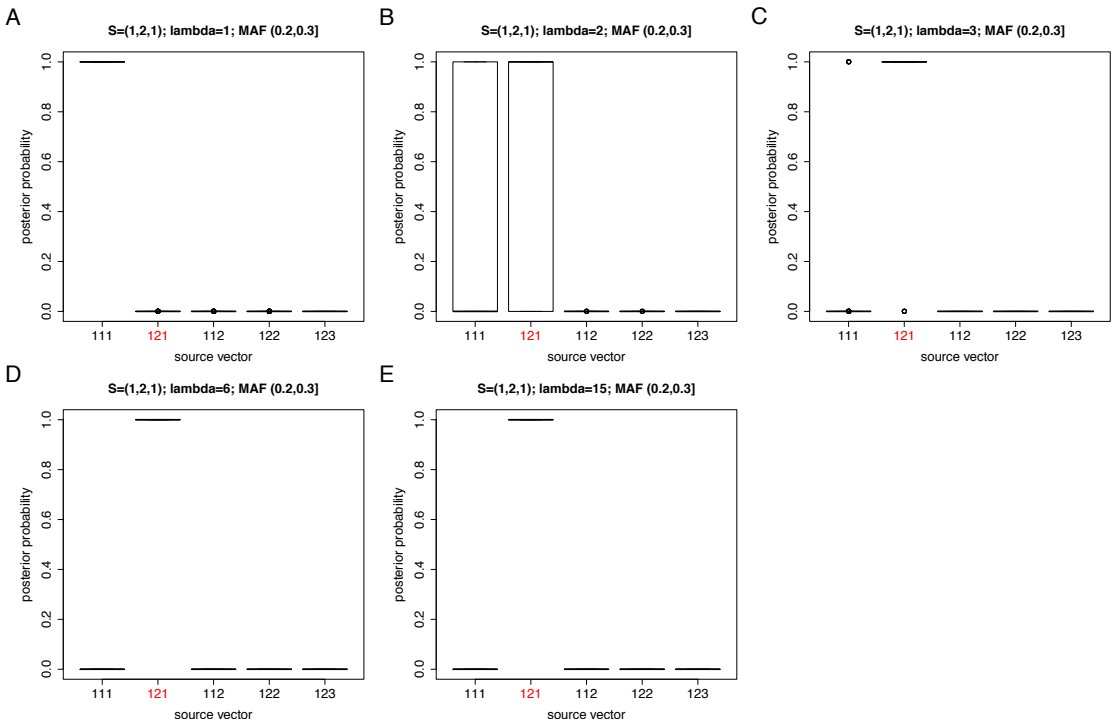
Appendix Figure 2.4 Additional plots for the simulation experiments outlined in "Simulation experiments to evaluate the impact of mean read depth and MAF on accuracy" for S = (1,2,1) and S = (1,2,3), showing the posterior probability assigned to all source vectors.

We present a series of four plots for the experiments where we simulated S = (1,2,1) and another four plots for the experiments where we simulated S = (1,2,3). Each plot consists of five subplots (one subplot for each of the tested mean read depths or lambdas). The title of each subplot shows the true (simulated) source vector for that experiment, the mean depth of putative replicates, and the MAF of sampled sites. Each subplot consists of five boxplots (one boxplot for each of the five possible source vectors). Each boxplot consists of 100 data points. We excluded plots for the (0.0,0.1] MAF interval since non zero probabilities were assigned only to S=(1,1,1) at every lambda. These plots reiterate the behavior observed in Figure 6 (main text) but do so at a higher resolution. For a given MAF interval, with the exception of (0.0,0.1], BIGRED shifts the probability away from S=(1,1,1) towards the true (simulated) source vector as the mean read depth of samples increases.

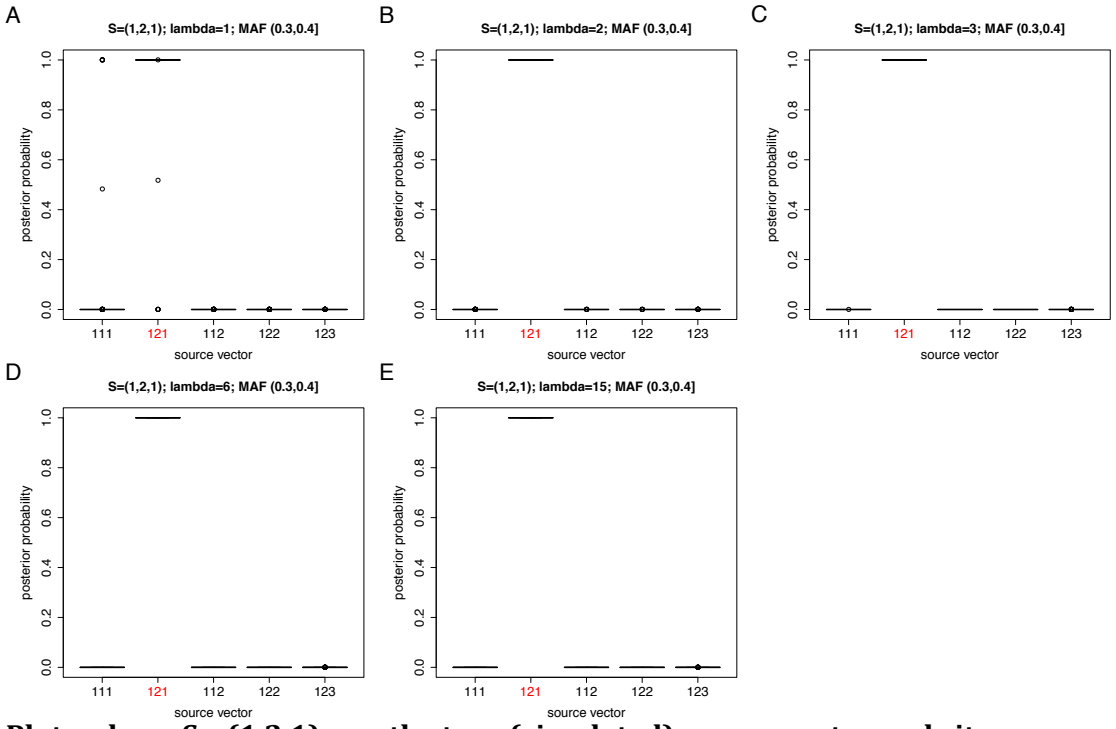
Plots for S = (1,2,1):



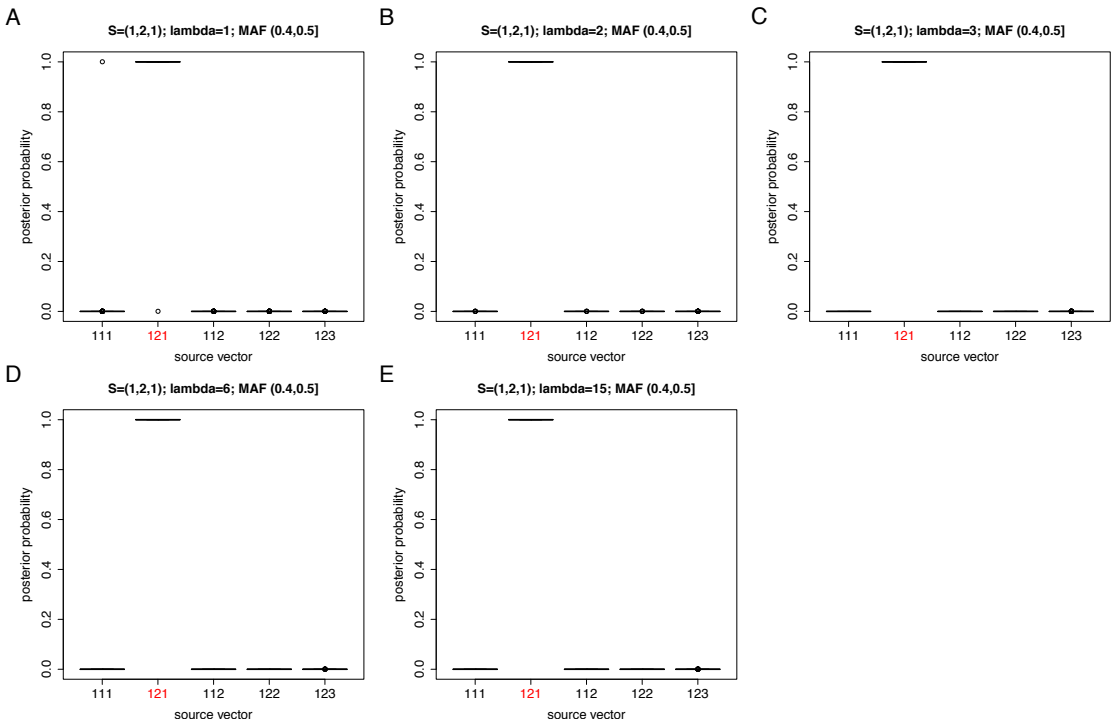
Plots where S = (1,2,1) was the true (simulated) source vector and sites were sampled from the (0.1,0.2] MAF interval.



Plots where *S* = (1,2,1) was the true (simulated) source vector and sites were sampled from the (0.2,0.3] MAF interval.

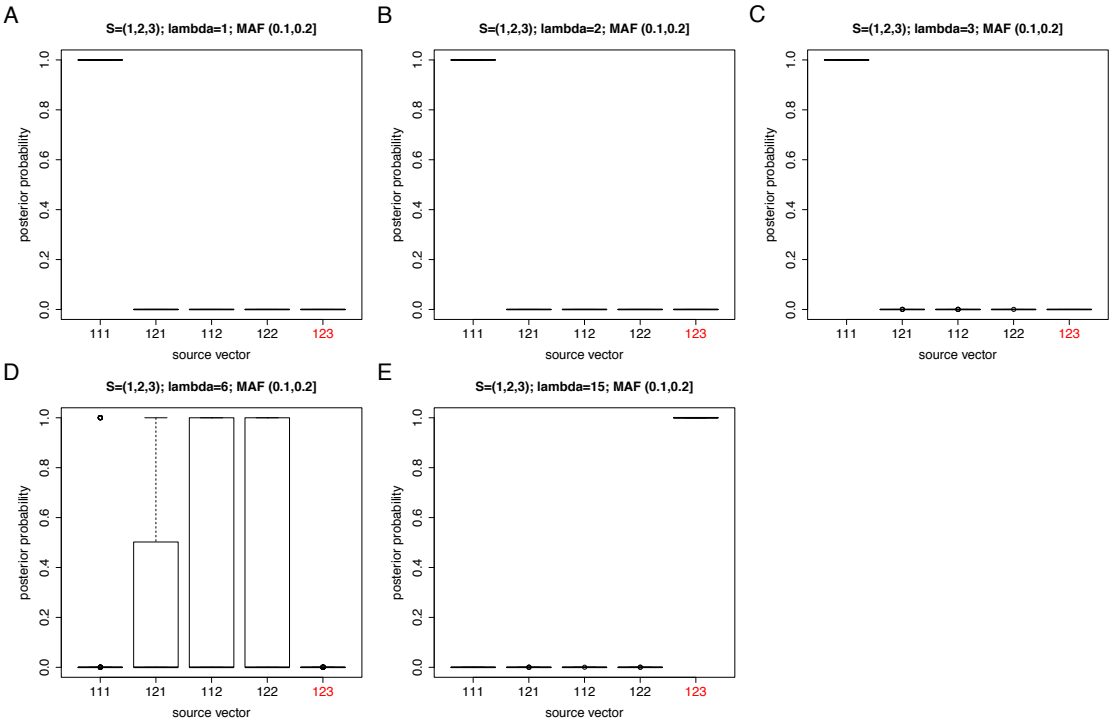


Plots where *S* = (1,2,1) was the true (simulated) source vector and sites were sampled from the (0.3,0.4] MAF interval.

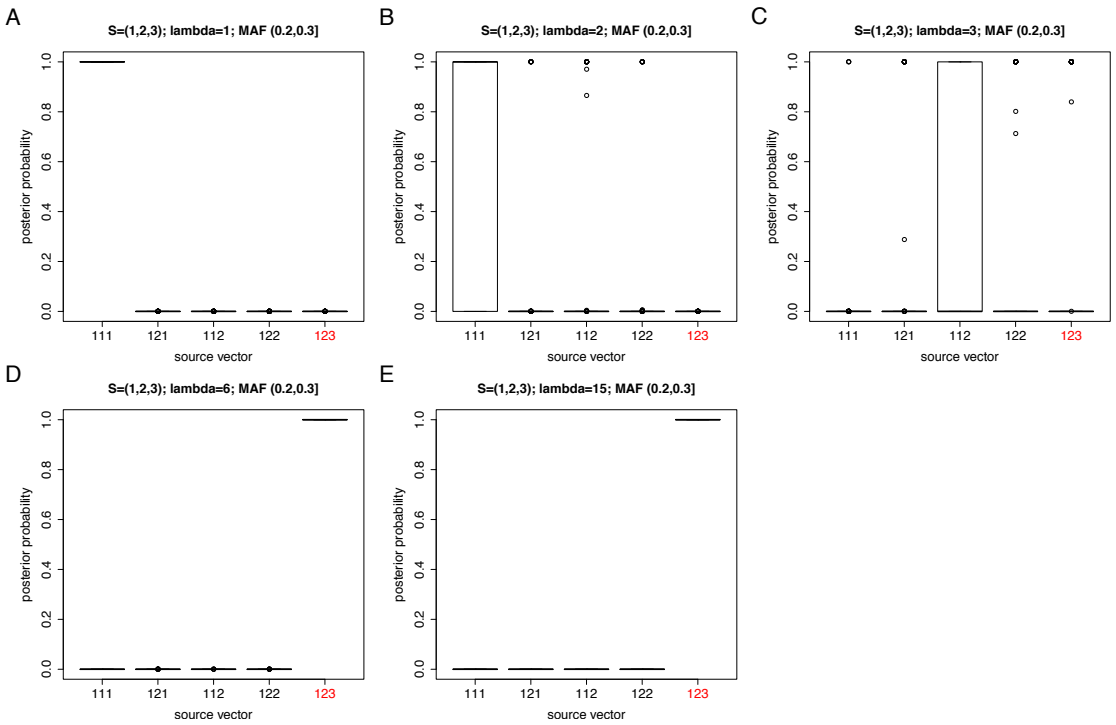


Plots where *S* = (1,2,1) was the true (simulated) source vector and sites were sampled from the (0.4,0.5] MAF interval.

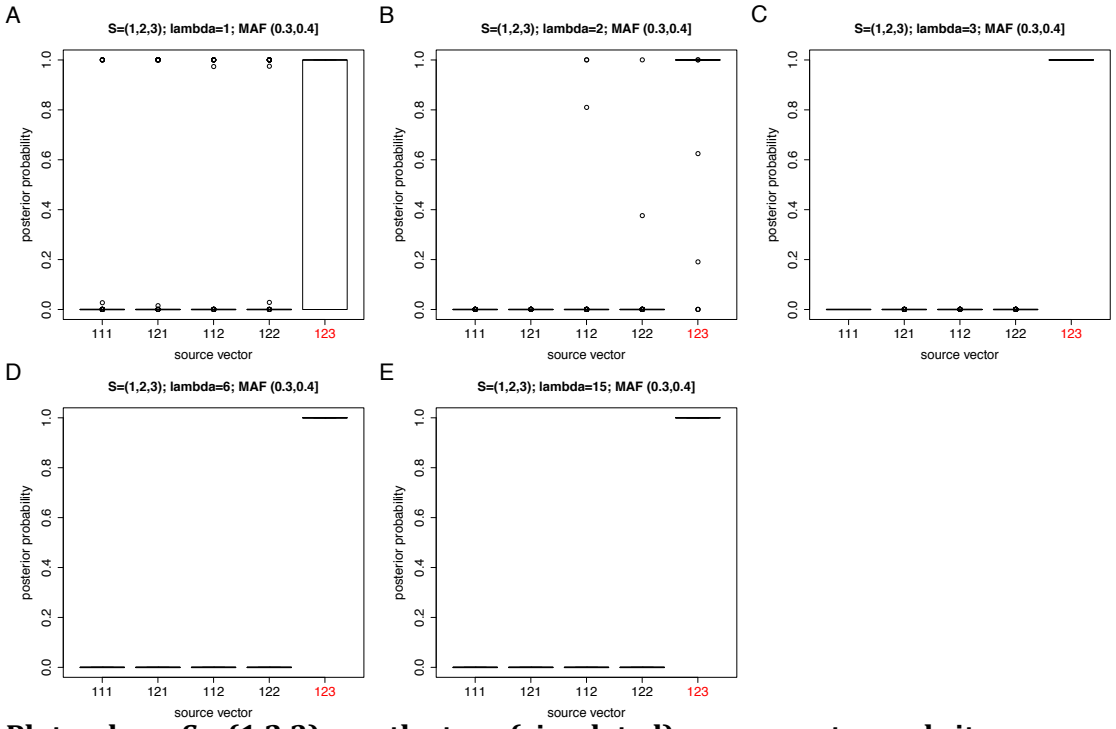
Plots for S = (1,2,3):



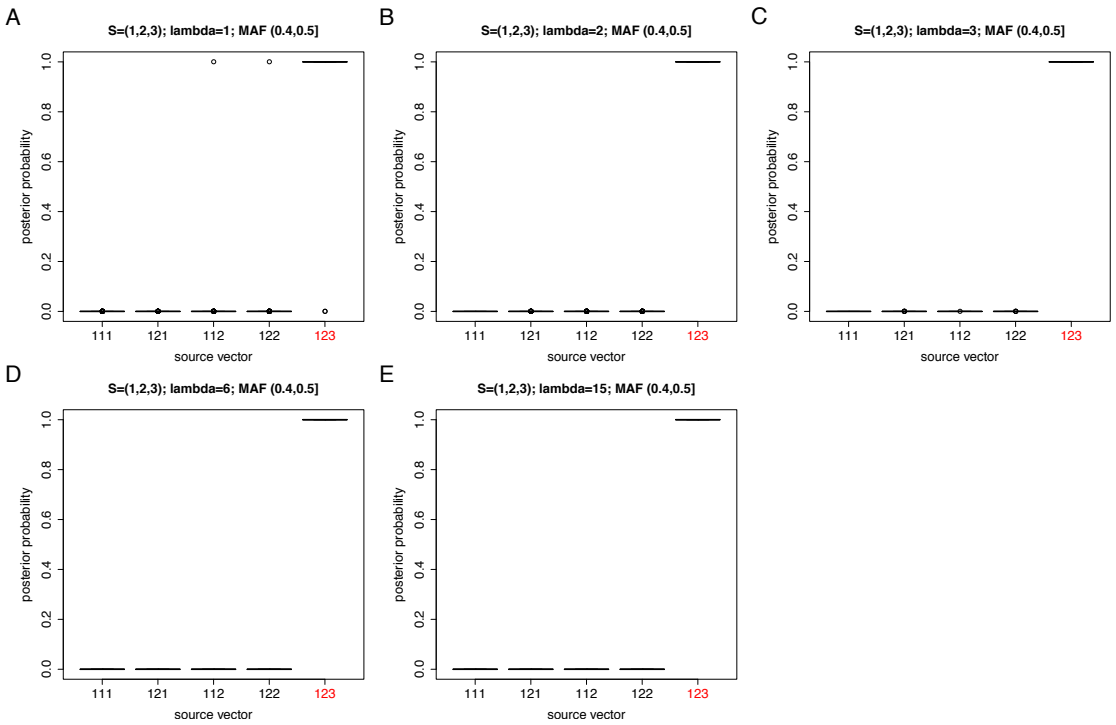
Plots where S = (1,2,3) was the true (simulated) source vector and sites were sampled from the (0.1,0.2] MAF interval.



Plots where *S* = (1,2,3) was the true (simulated) source vector and sites were sampled from the (0.2,0.3] MAF interval.

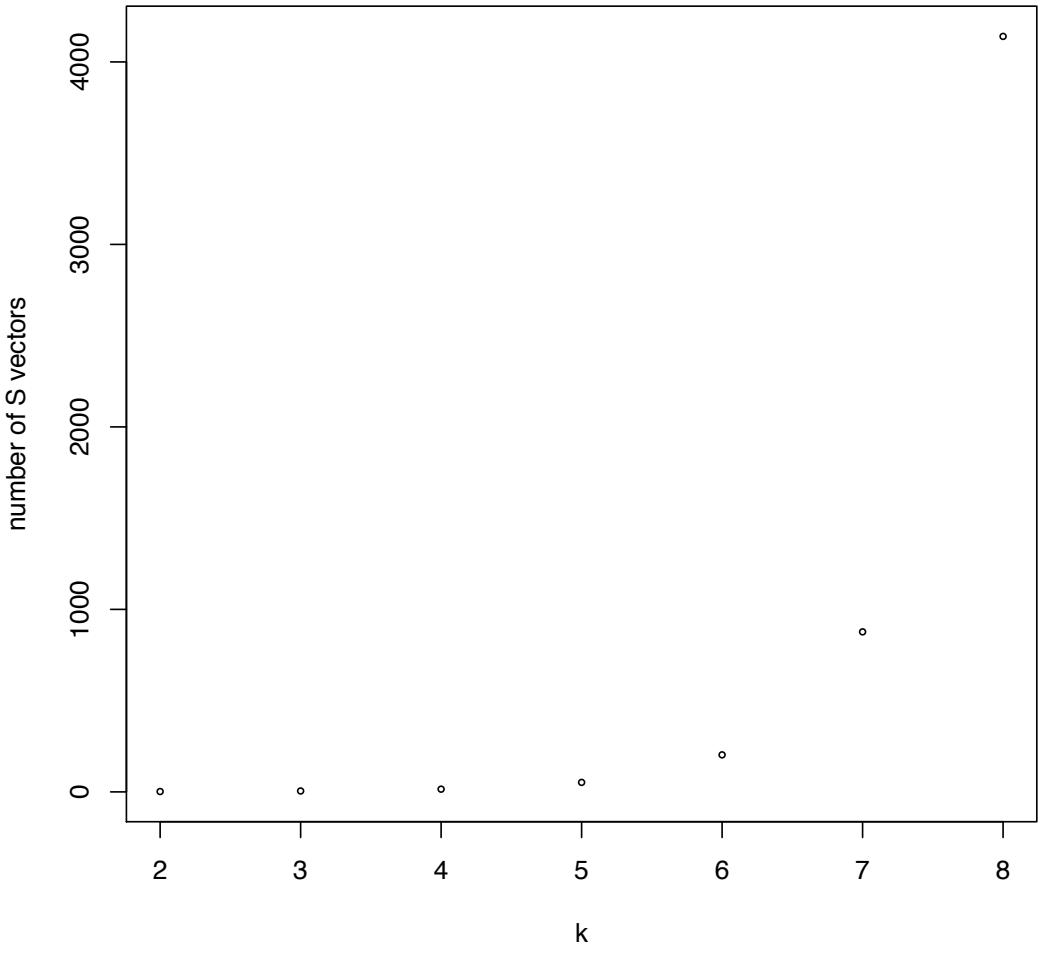


Plots where *S* = (1,2,3) was the true (simulated) source vector and sites were sampled from the (0.3,0.4] MAF interval.



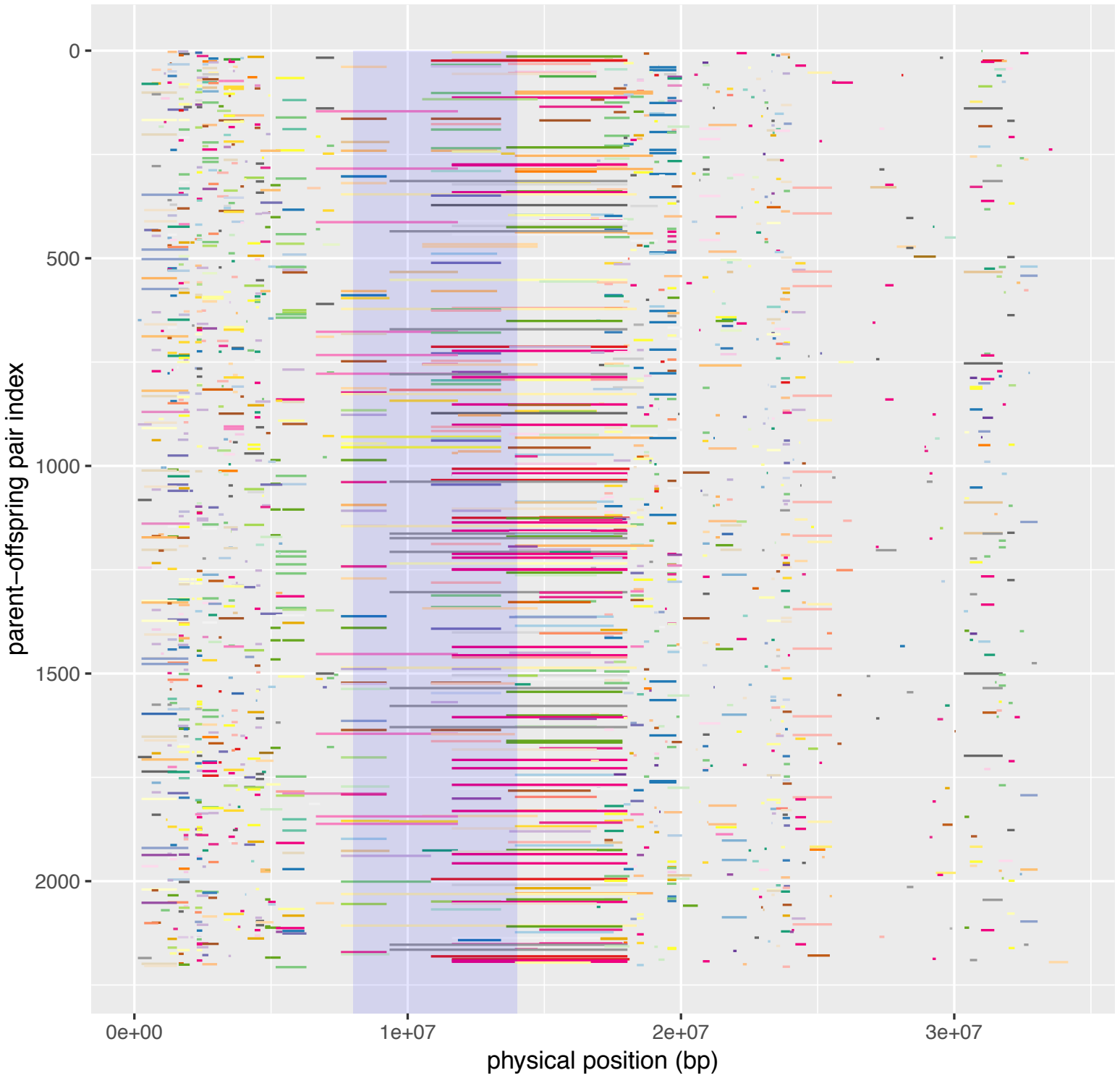
Plots where *S* = (1,2,3) was the true (simulated) source vector and sites were sampled from the (0.4,0.5] MAF interval.

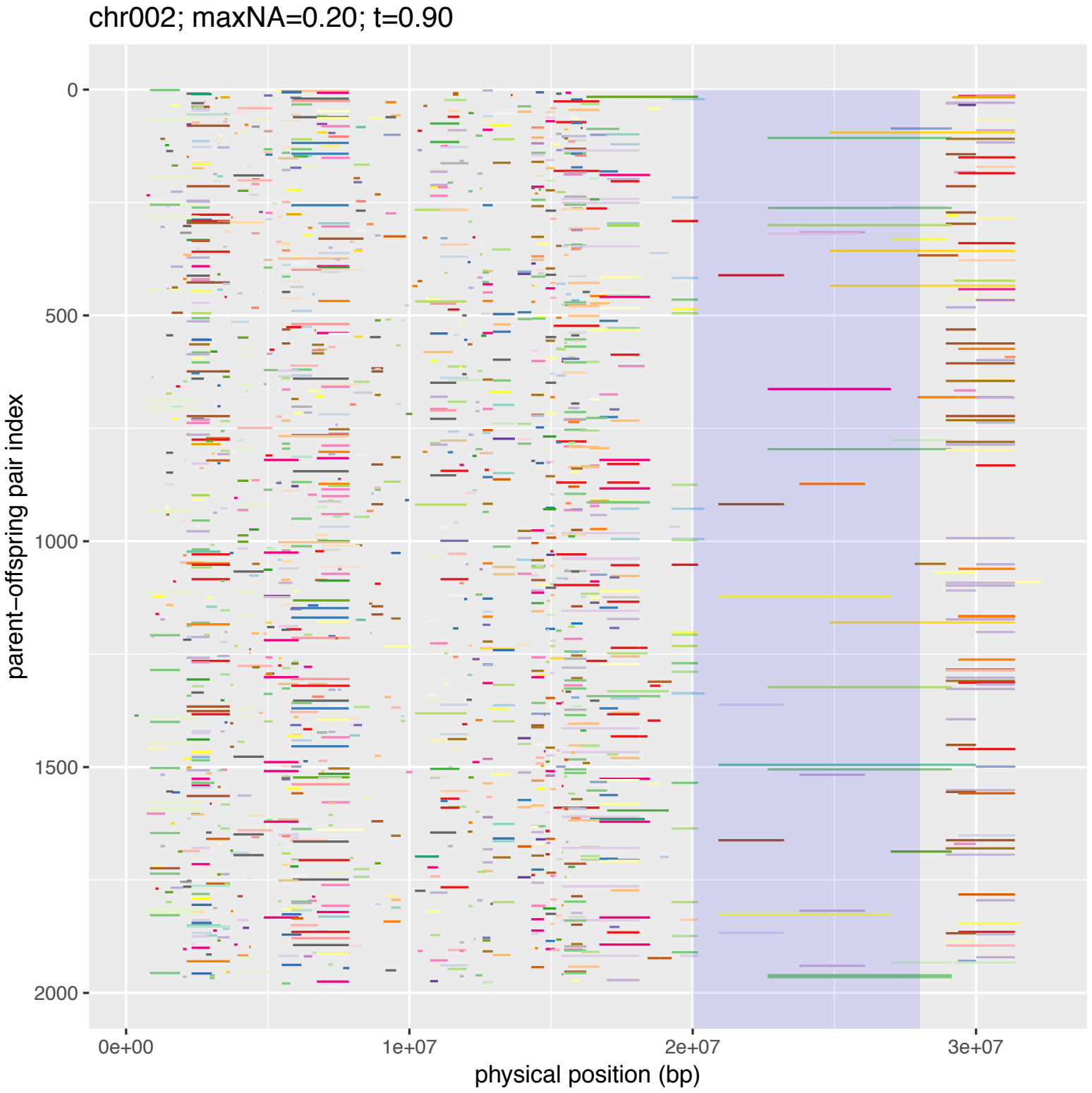
Appendix Figure 2.5 Plot showing the number of source vectors associated with k for $k \in \{1, \ldots, 8\}$.

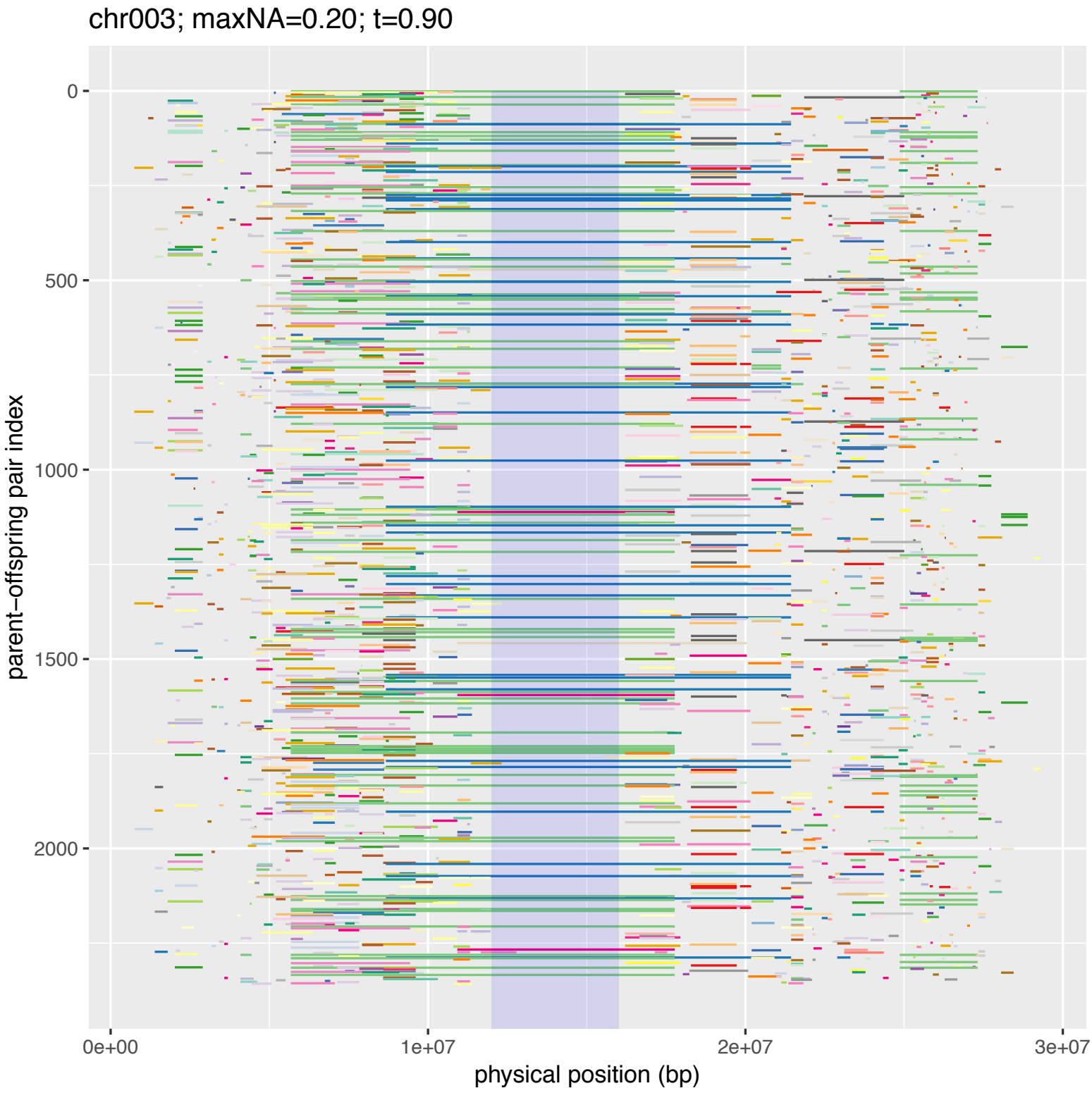


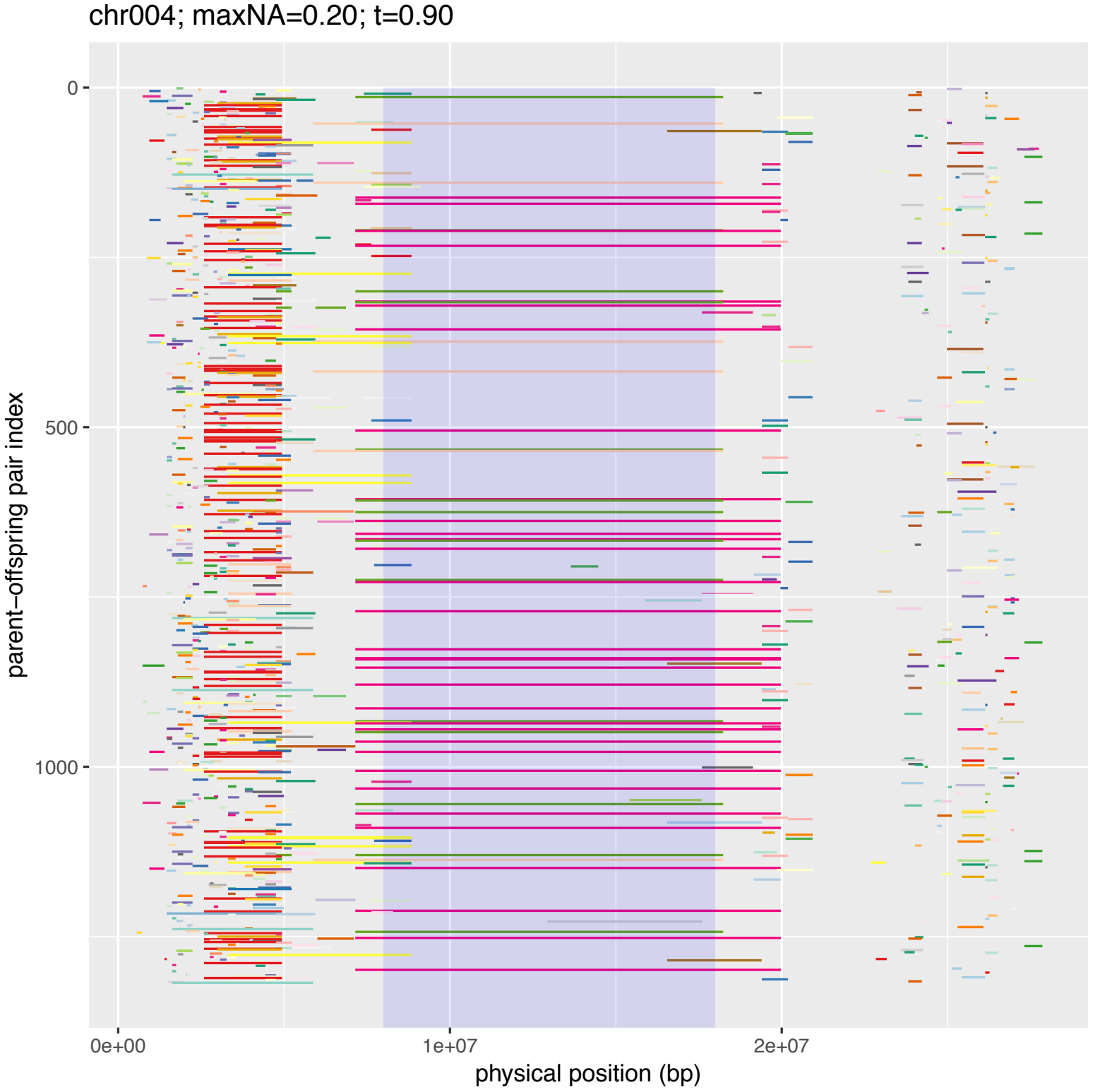
Appendix Figure 3.1

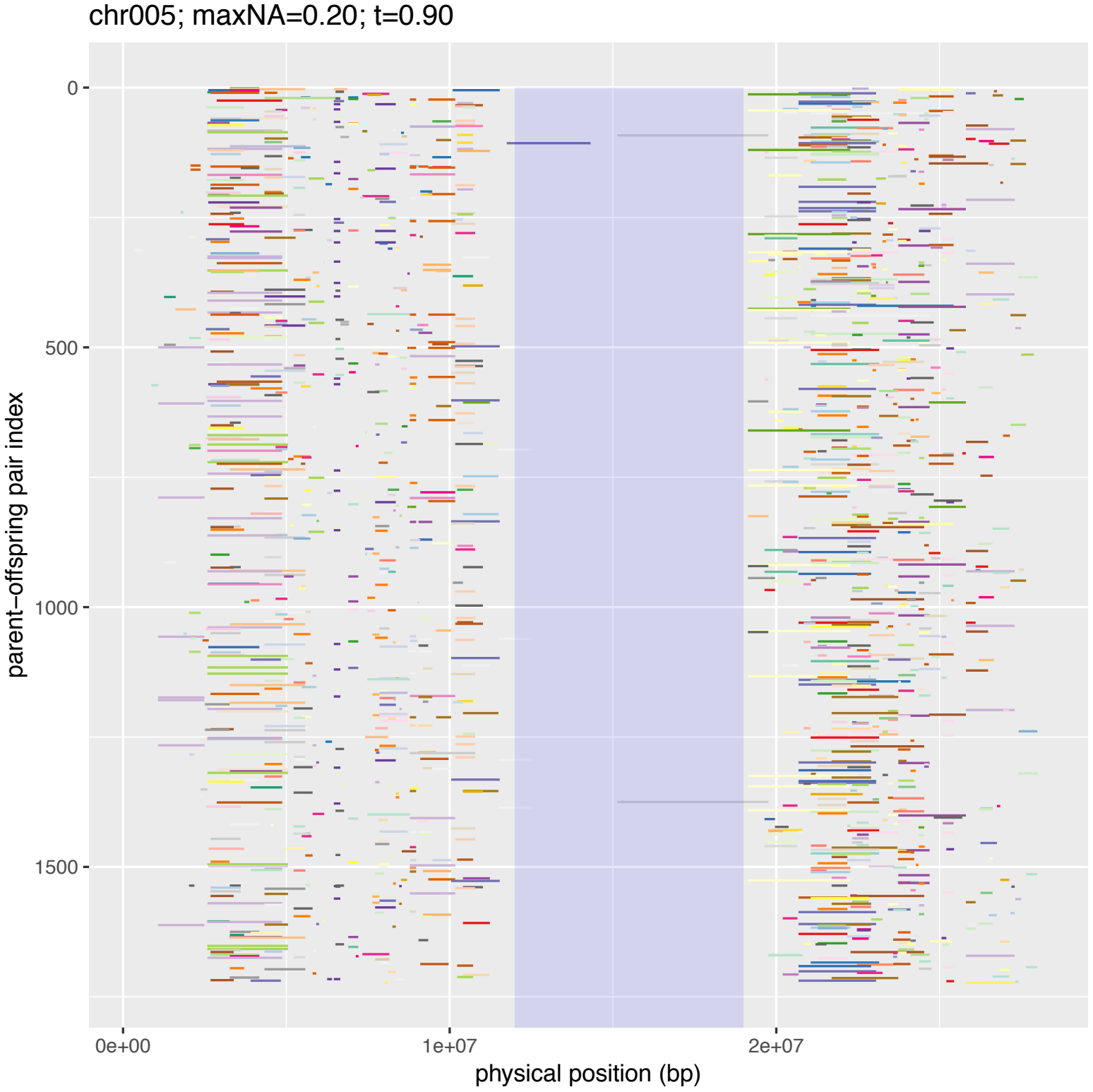
The first set of 18 plots in this figure show SHAPEIT2-duoHMM results when using the 20% maximum-missing filter. The second set of 18 plots show the results when using the 30% maximum-missing filter. The x-axis of each plot shows the physical position (bp) of each chromosome. Each colored, horizontal line in the plot represents a region where a crossover has occurred (with probability $t = 0.9$). Each parent has been assigned a unique color. Crossover intervals detected in the same parent-offspring duo appear on the same row and in the same color. The centromere for each chromosome is shaded blue. The title of each plot shows the chromosome, maximum-missing threshold, and t threshold.

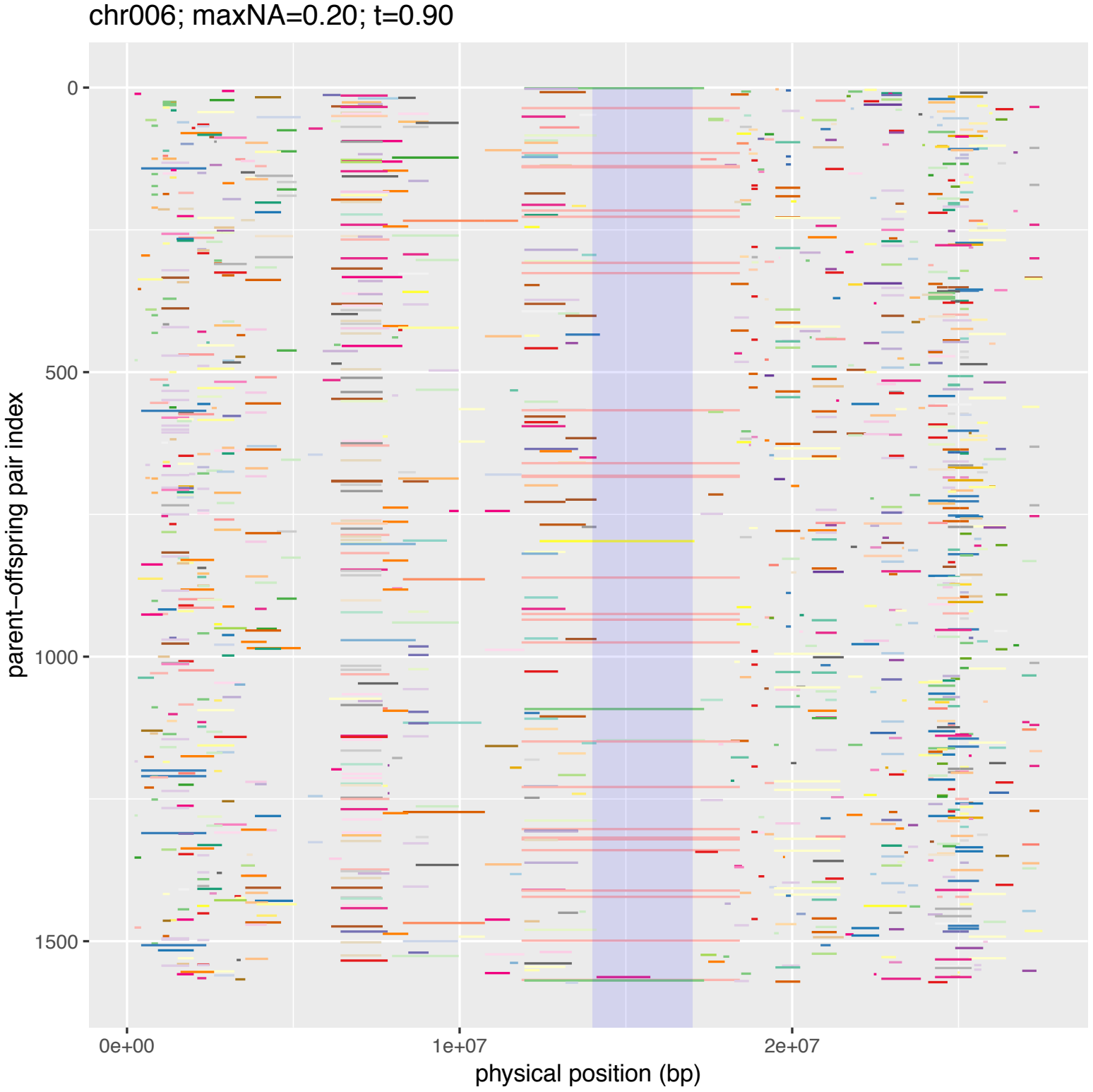

chr001; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
1500
2000
0e+00
1e+07
2e+07
3e+07
physical position (bp)

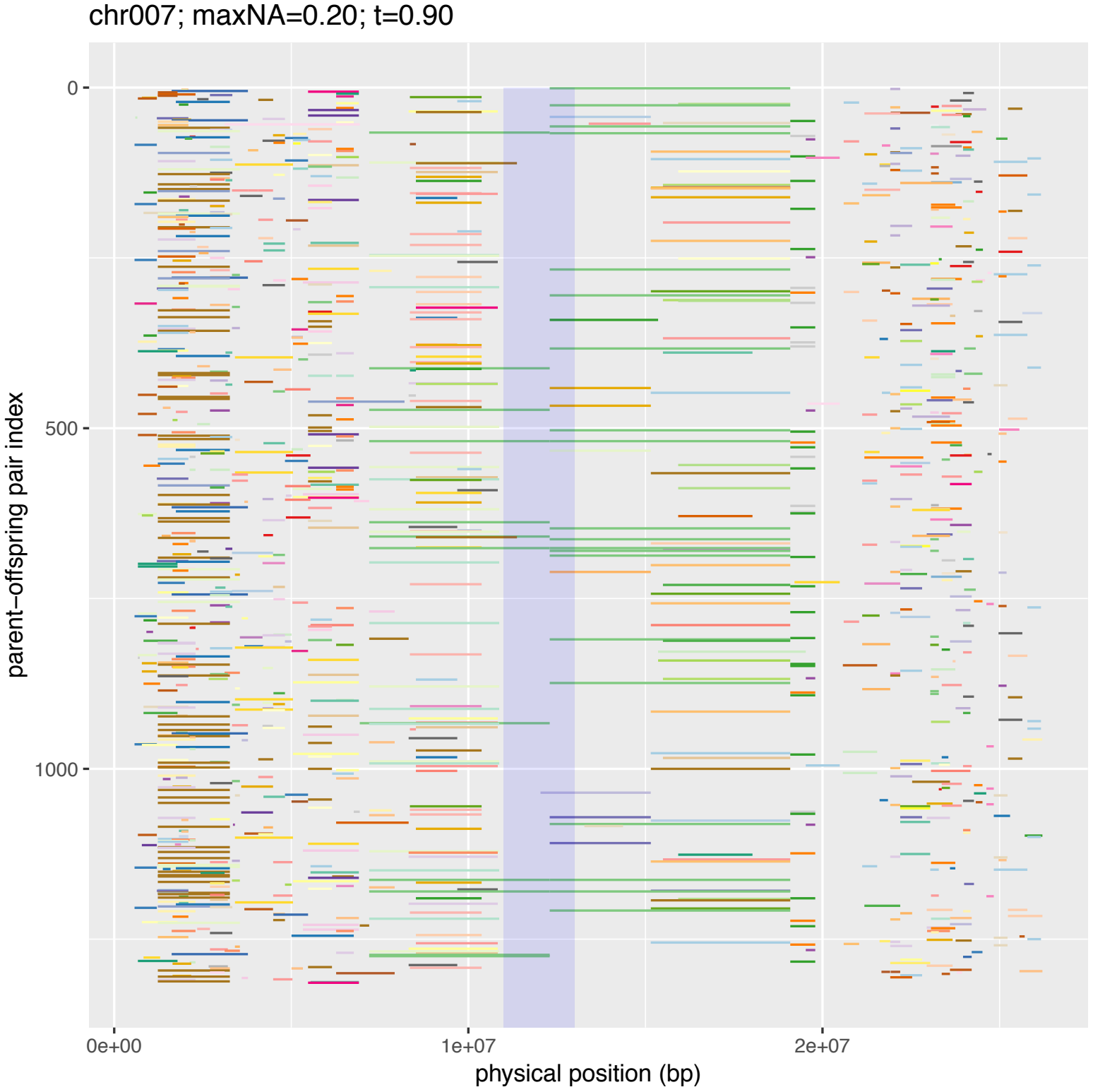

chr002; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
1500
2000
0e+00
1e+07
2e+07
3e+07
physical position (bp)

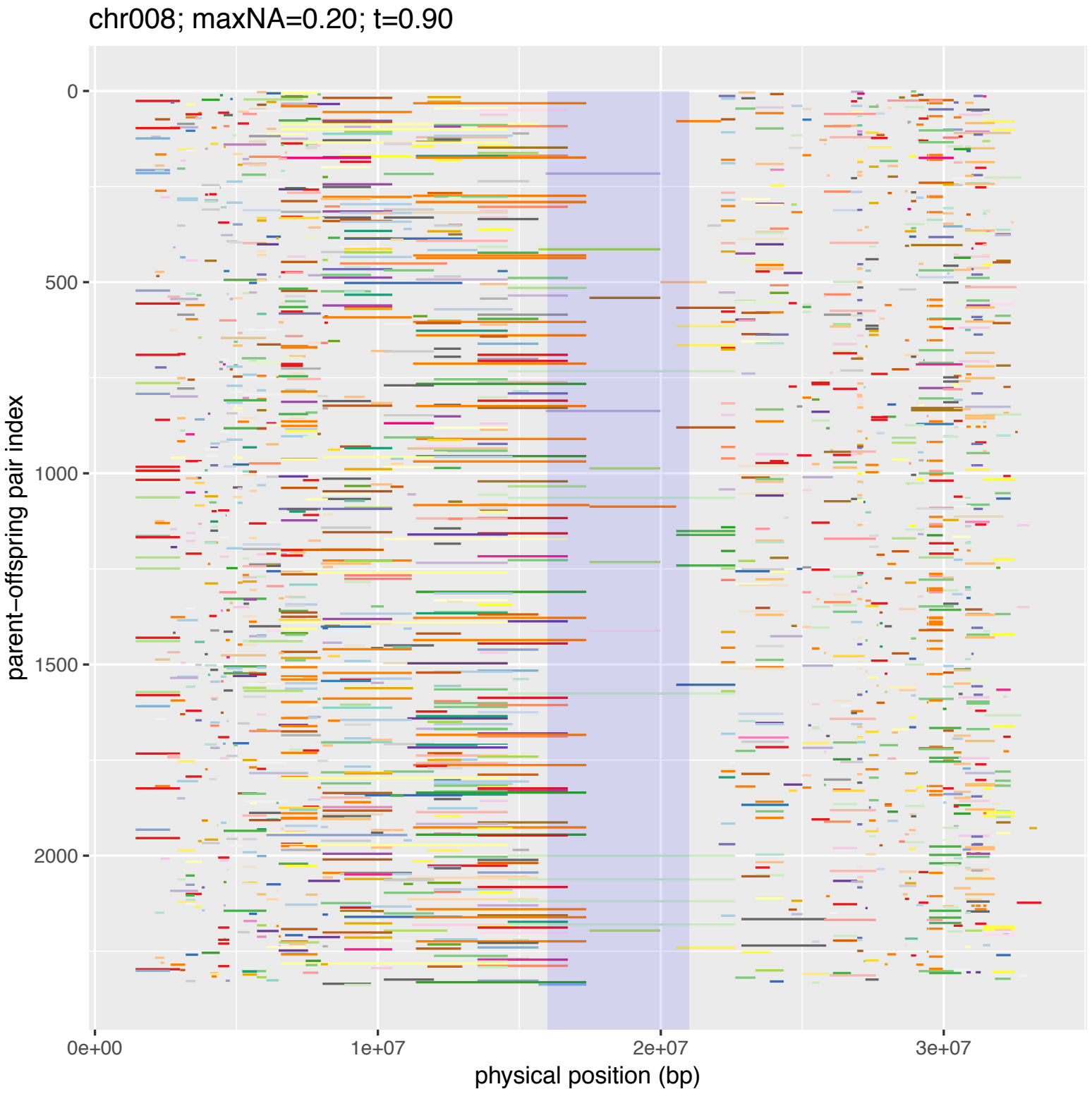

chr003; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
1500
2000
0e+00
1e+07
2e+07
3e+07
physical position (bp)

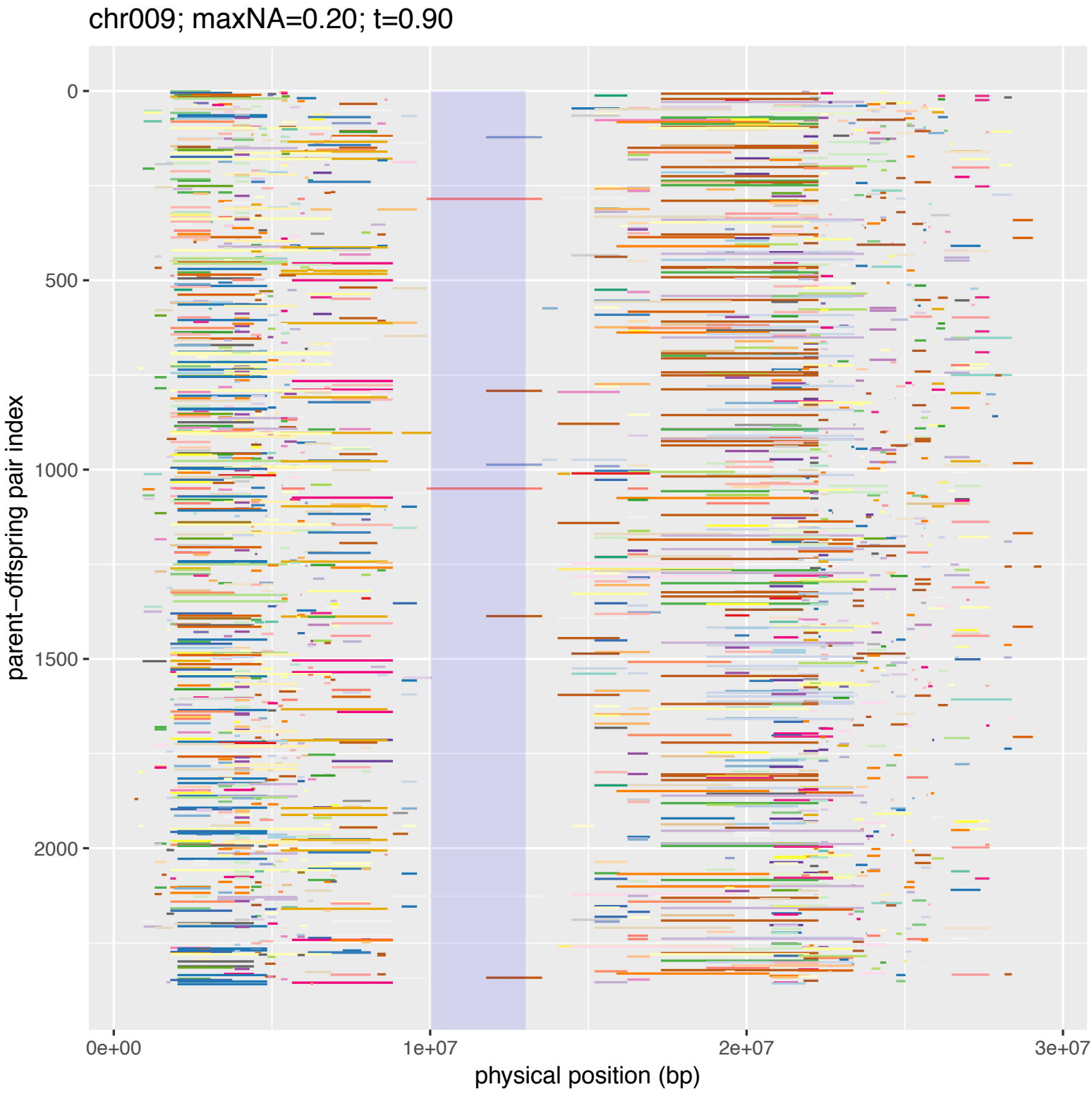

chr004; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
0e+00
1e+07
2e+07
physical position (bp)

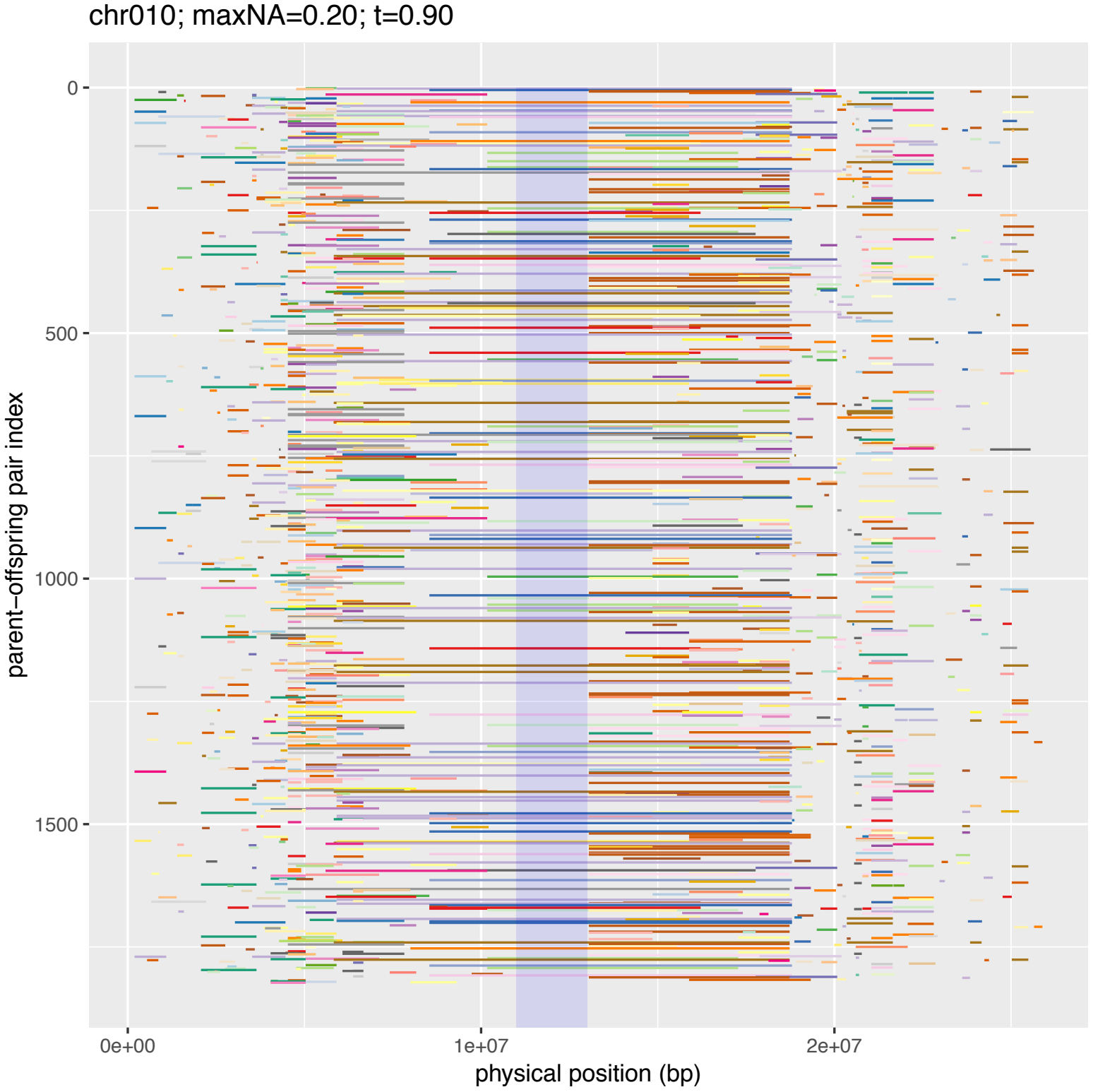

chr005; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
1500
0e+00
1e+07
2e+07
physical position (bp)

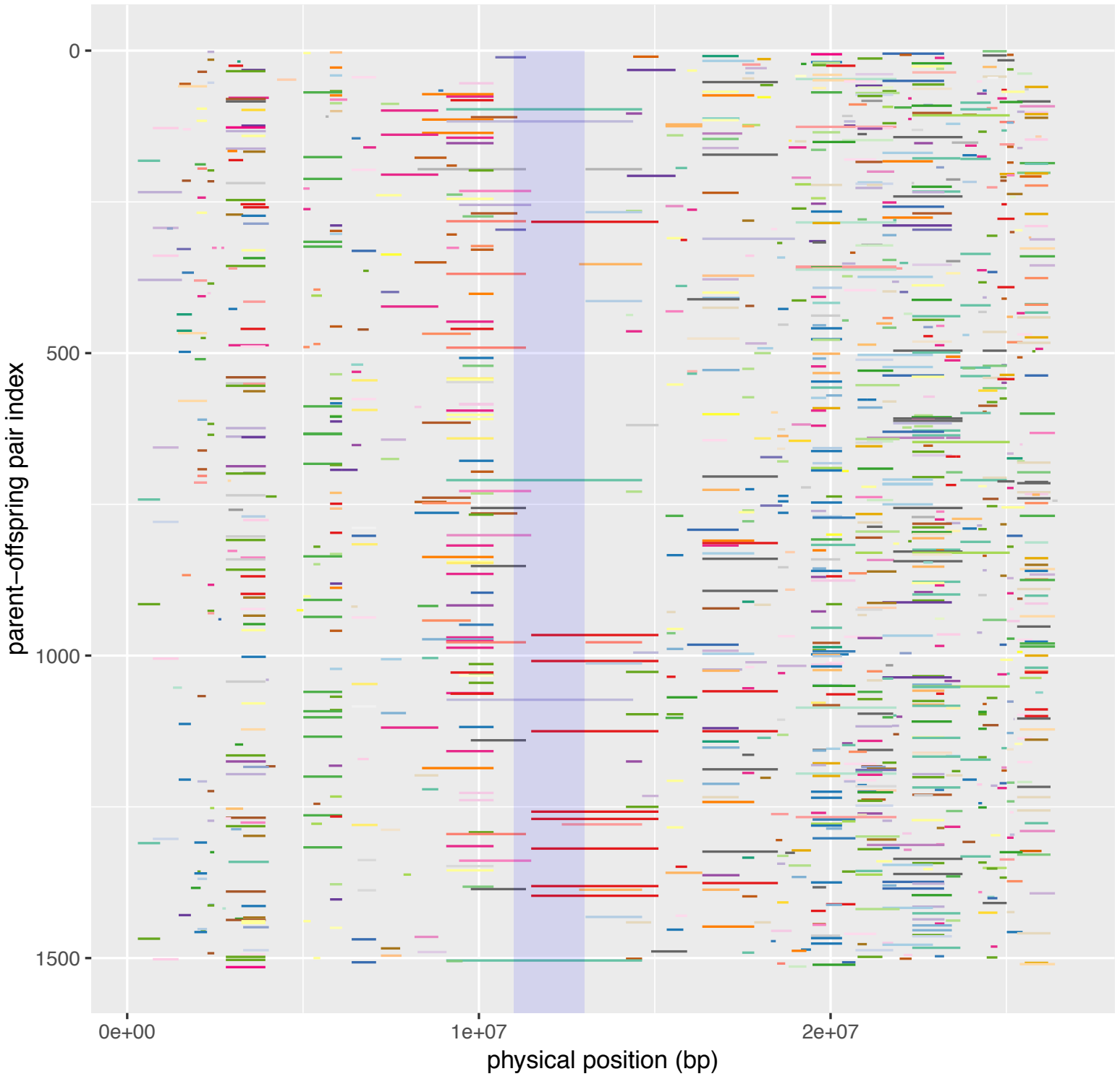

chr006; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
1500
0e+00
1e+07
2e+07
physical position (bp)

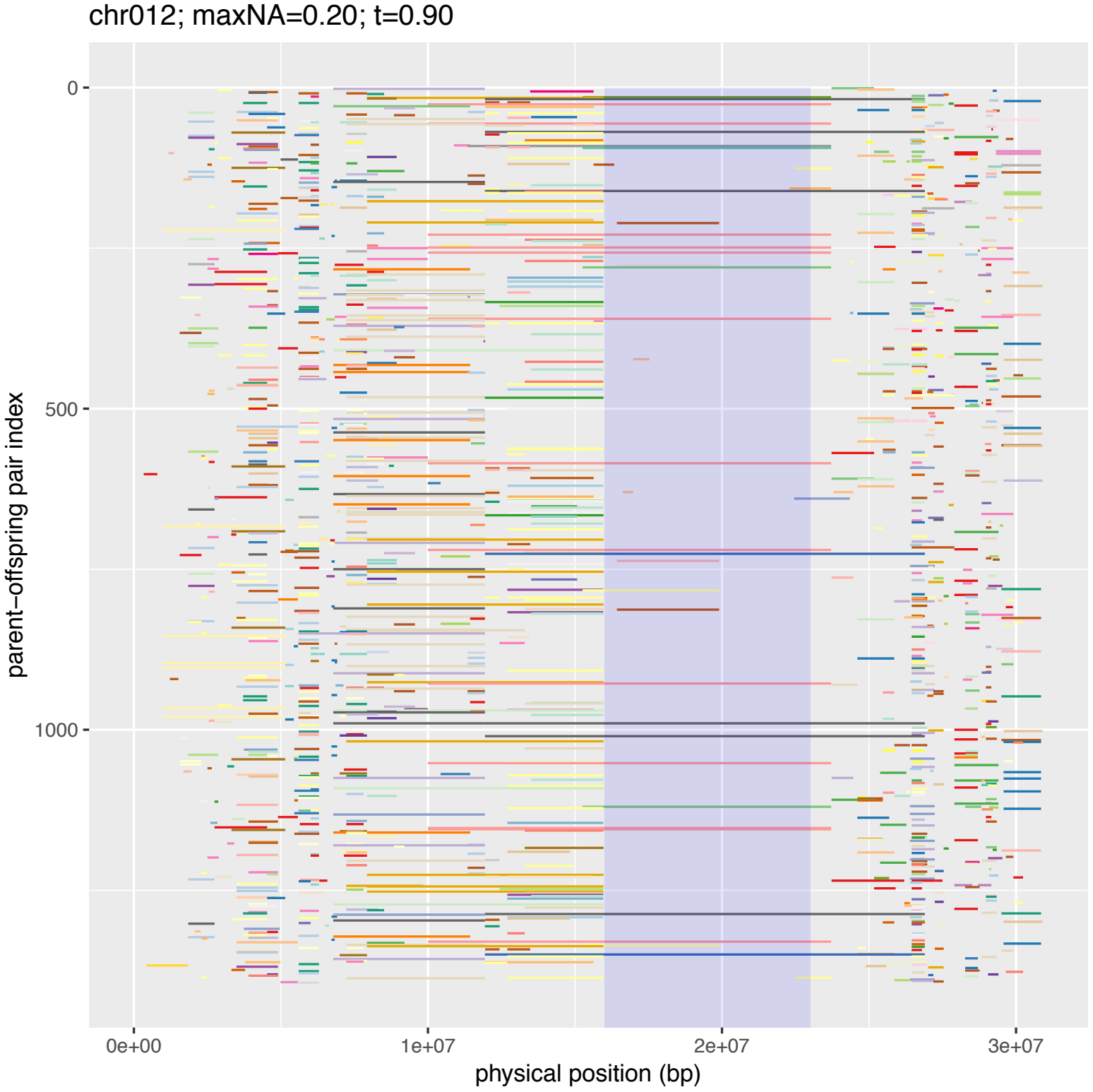

chr007; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
0e+00
1e+07
2e+07
physical position (bp)

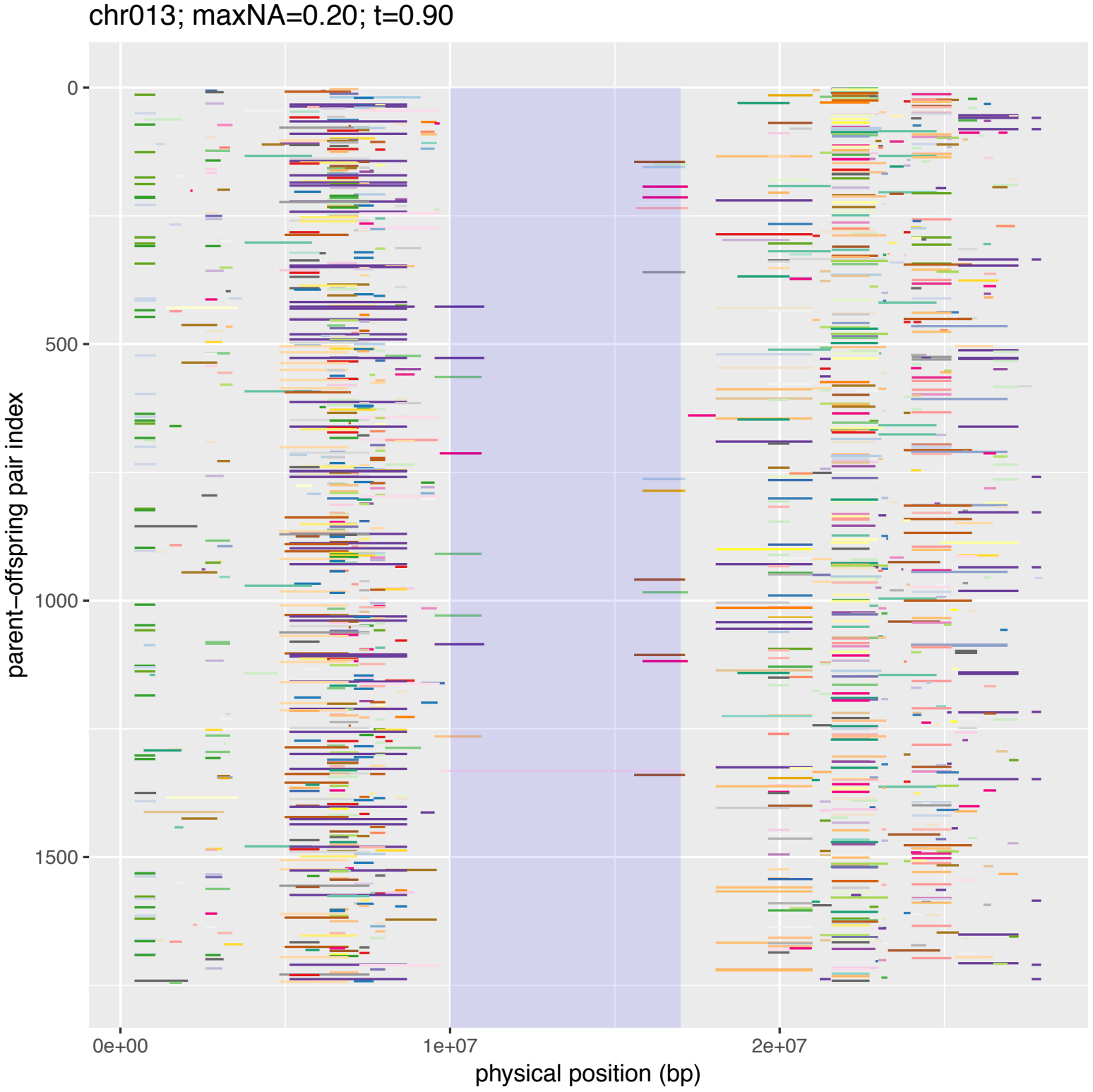

chr008; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
1500
2000
0e+00
1e+07
2e+07
3e+07
physical position (bp)

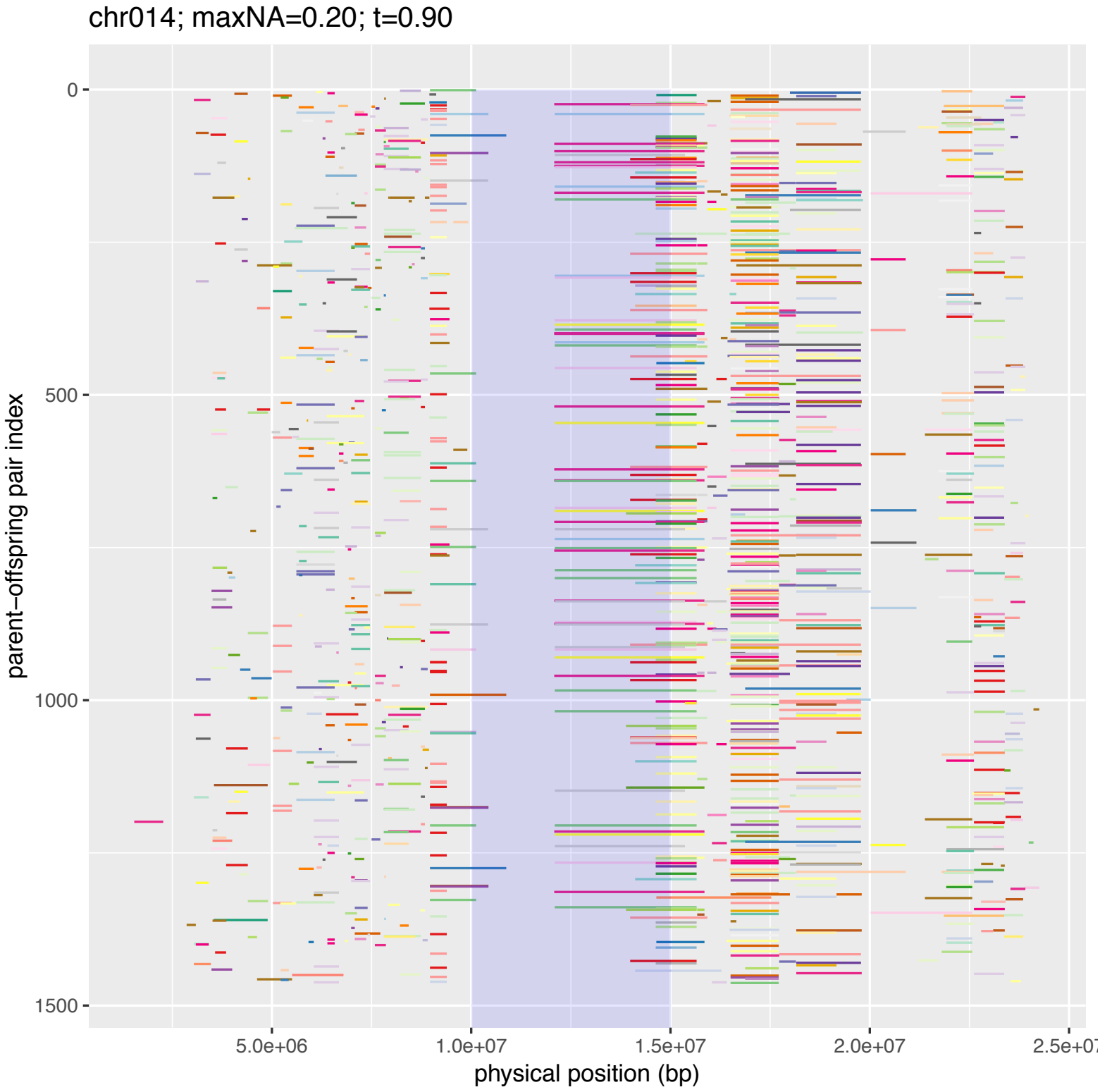

chr009; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
1500
2000
0e+00
1e+07
2e+07
3e+07
physical position (bp)

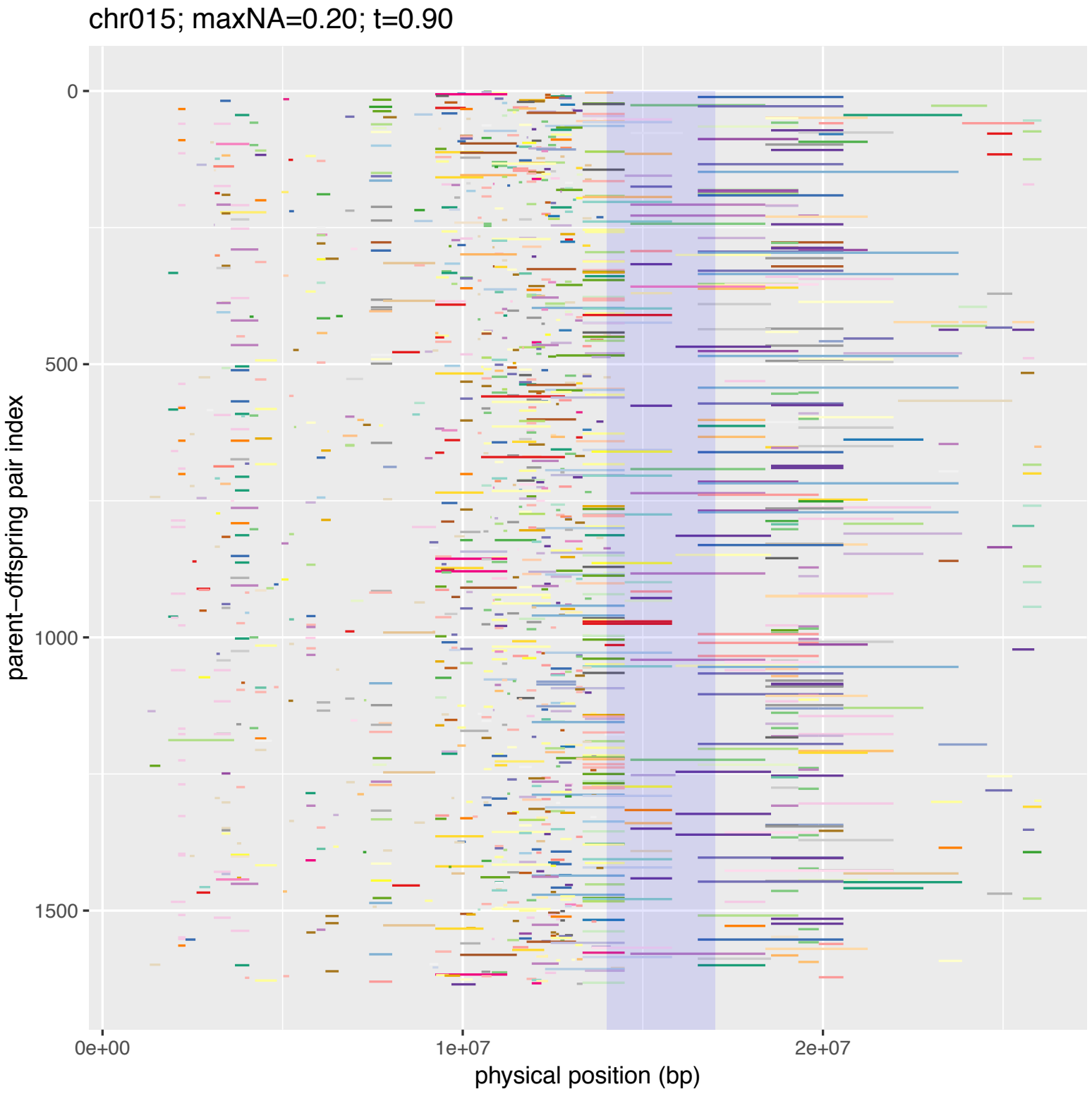

chr010; maxNA=0.20; t=0.90
0
500
1000
1500
parent−offspring pair index
0e+00
1e+07
2e+07
physical position (bp)

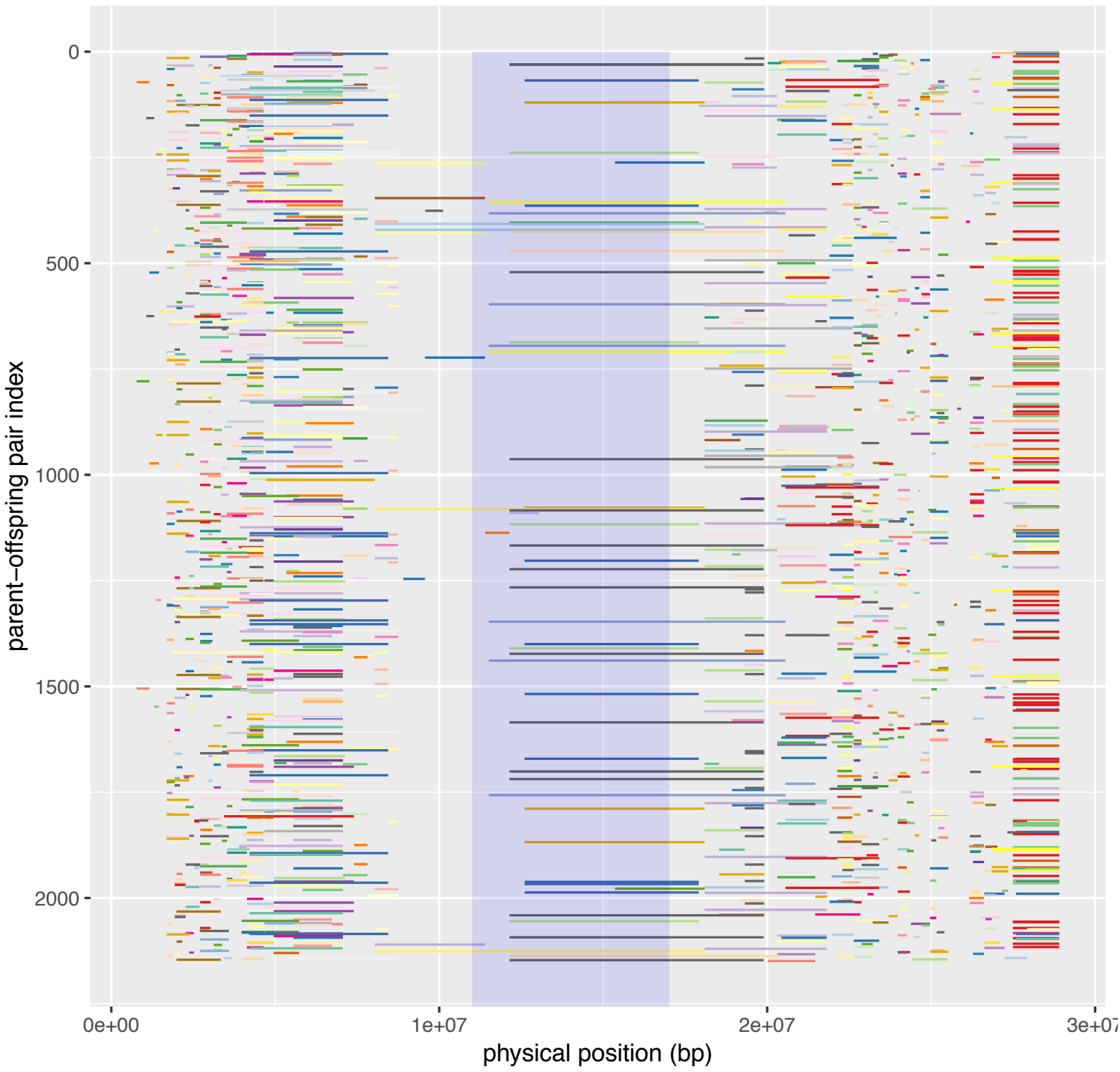

chr011; maxNA=0.20; t=0.90
0
500
1000
1500
parent−offspring pair index
0e+00
1e+07
2e+07
physical position (bp)

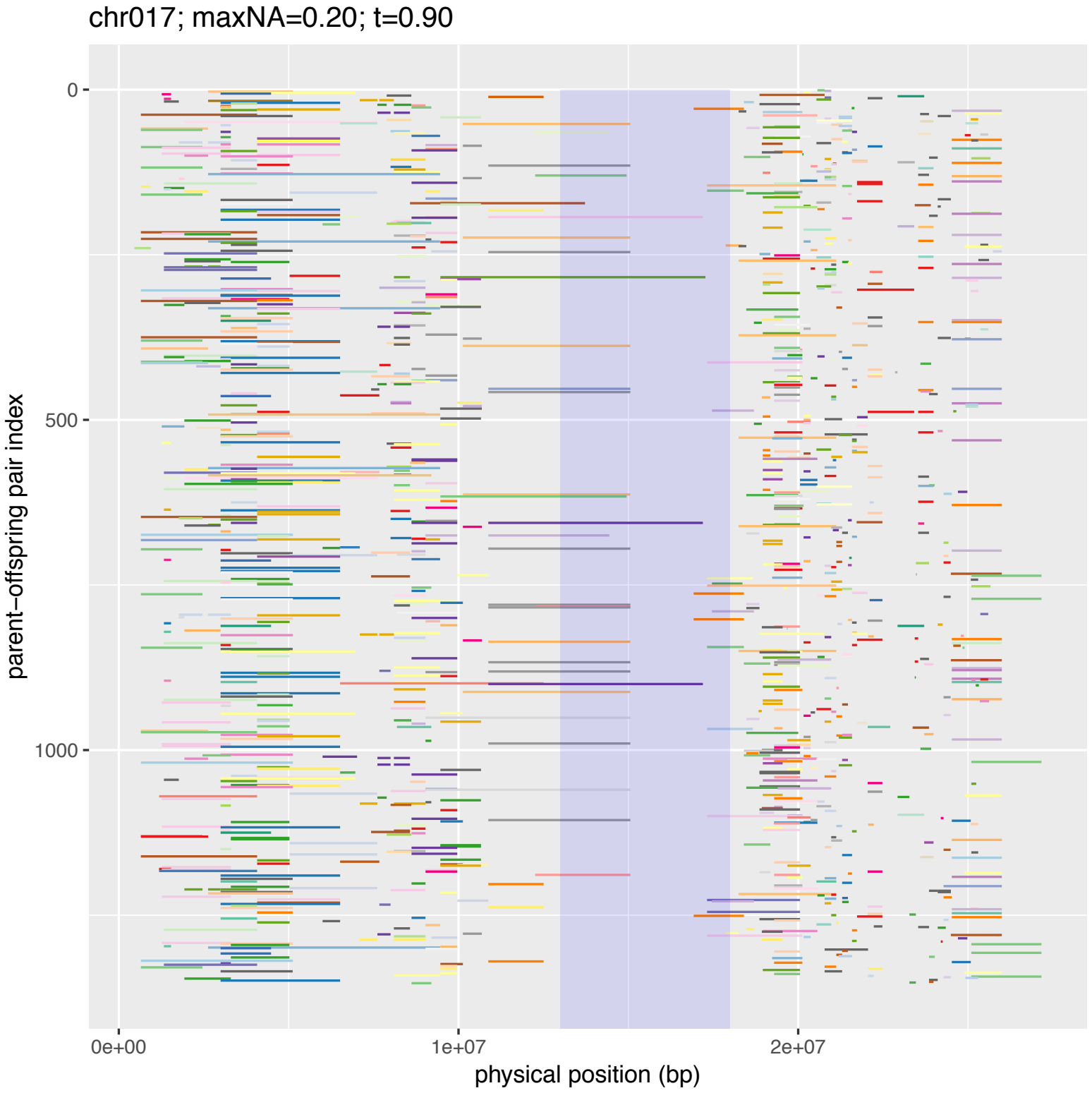

chr012; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
0e+00
1e+07
2e+07
3e+07
physical position (bp)

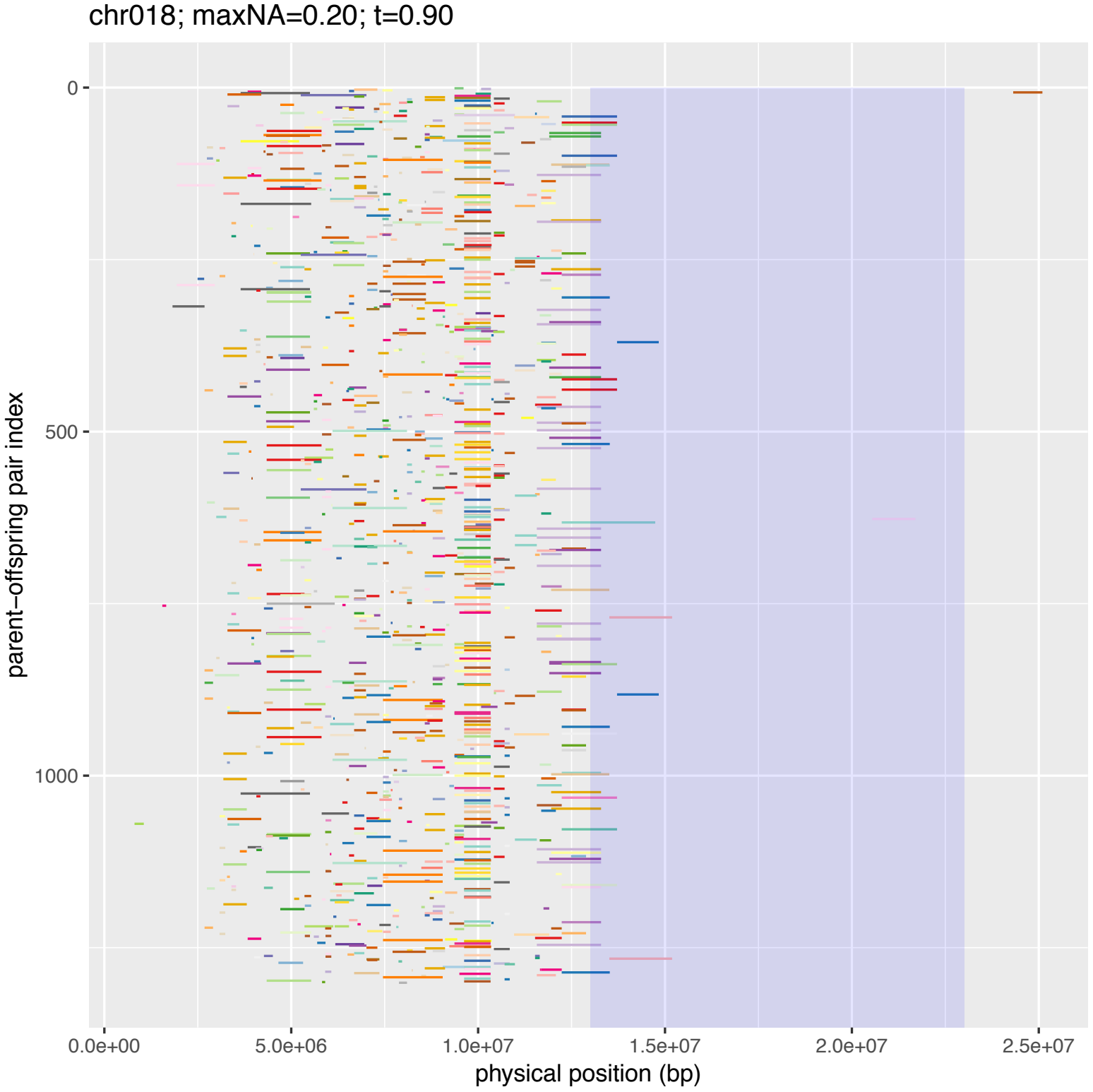

chr013; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
1500
0e+00
1e+07
2e+07
physical position (bp)

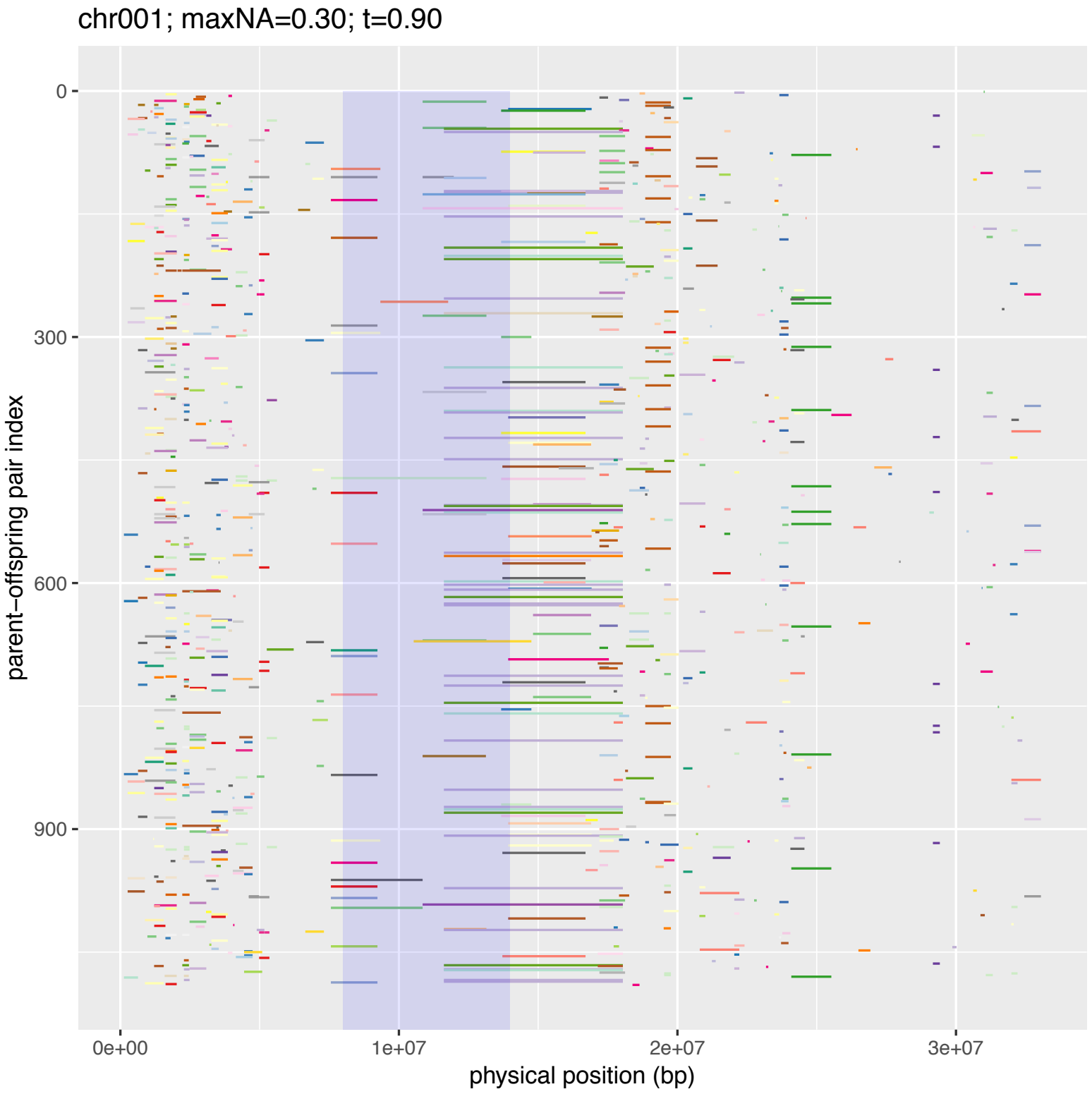

chr014; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
1500
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
physical position (bp)

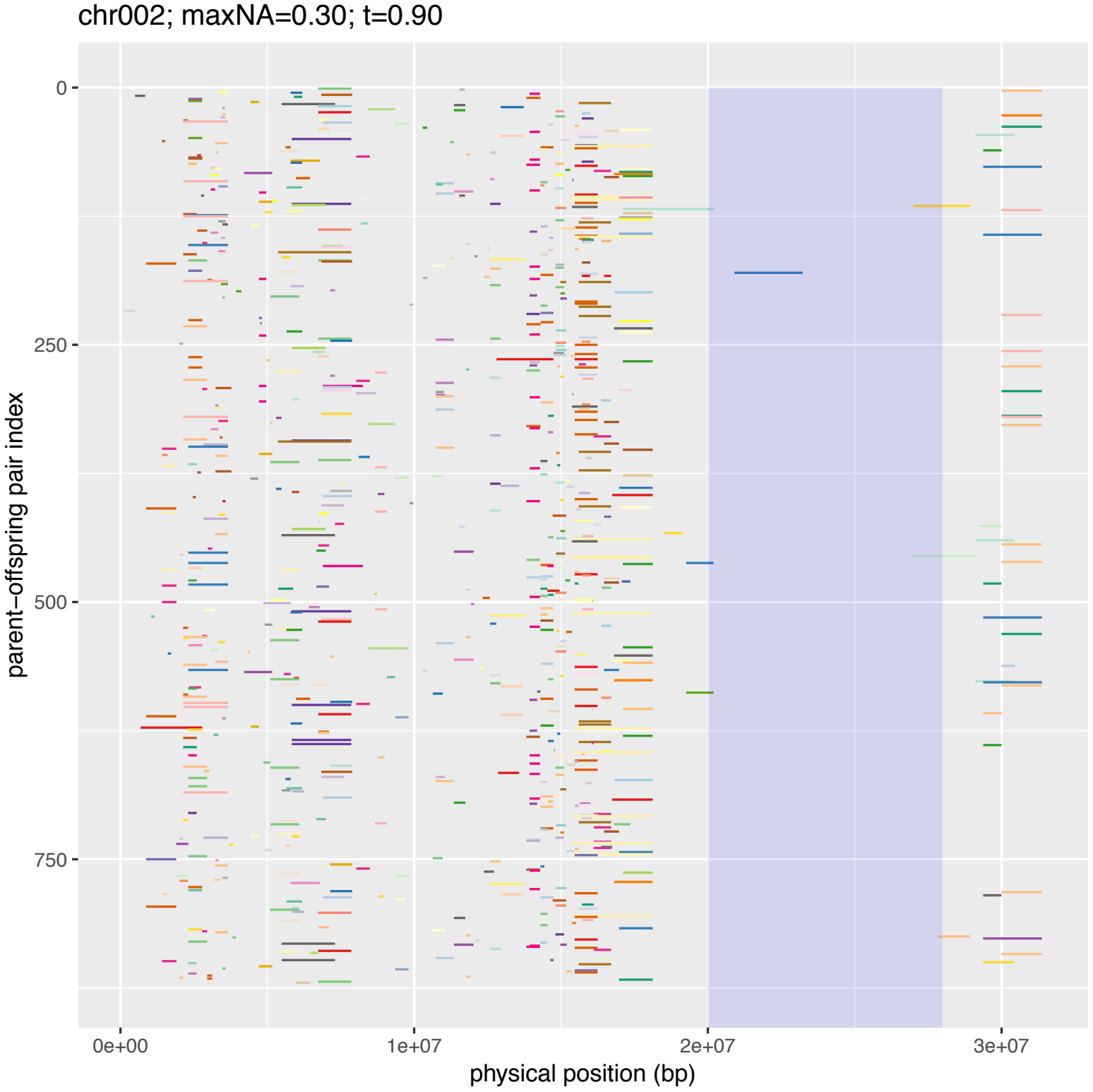

chr015; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
1500
0e+00
1e+07
2e+07
physical position (bp)

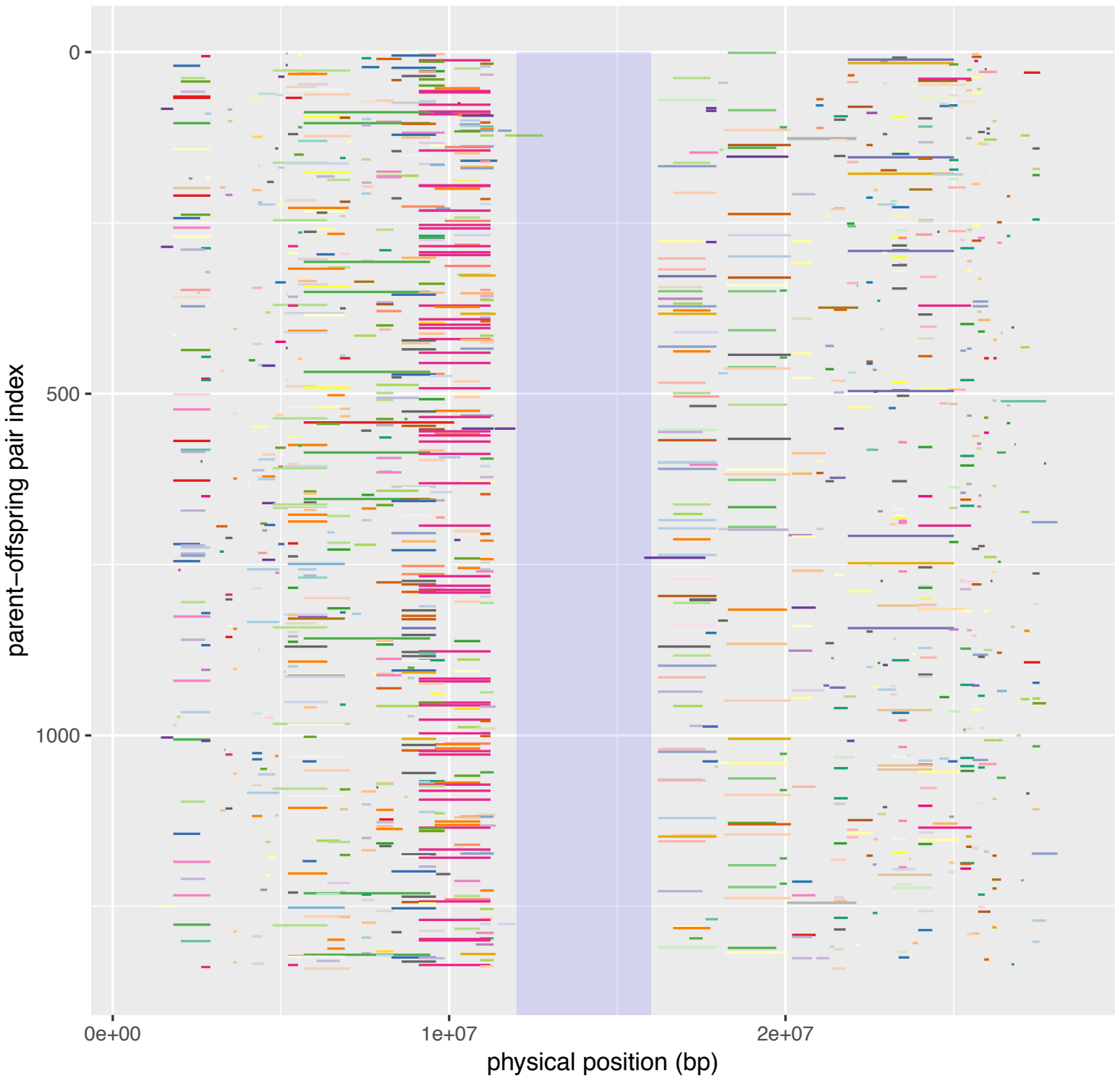

chr016; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
1500
2000
0e+00
1e+07
2e+07
3e+07
physical position (bp)

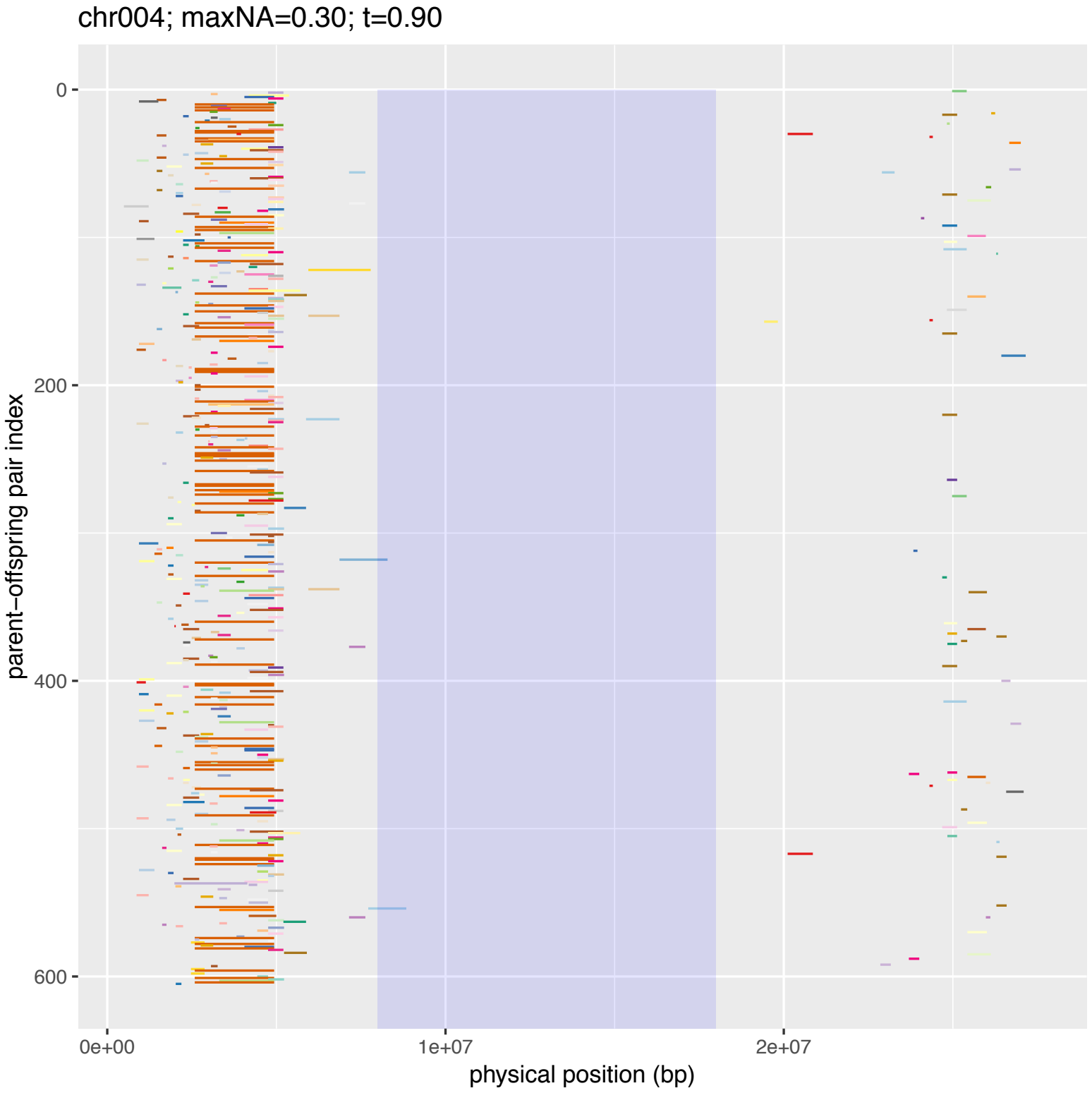

chr017; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
0e+00
1e+07
2e+07
physical position (bp)

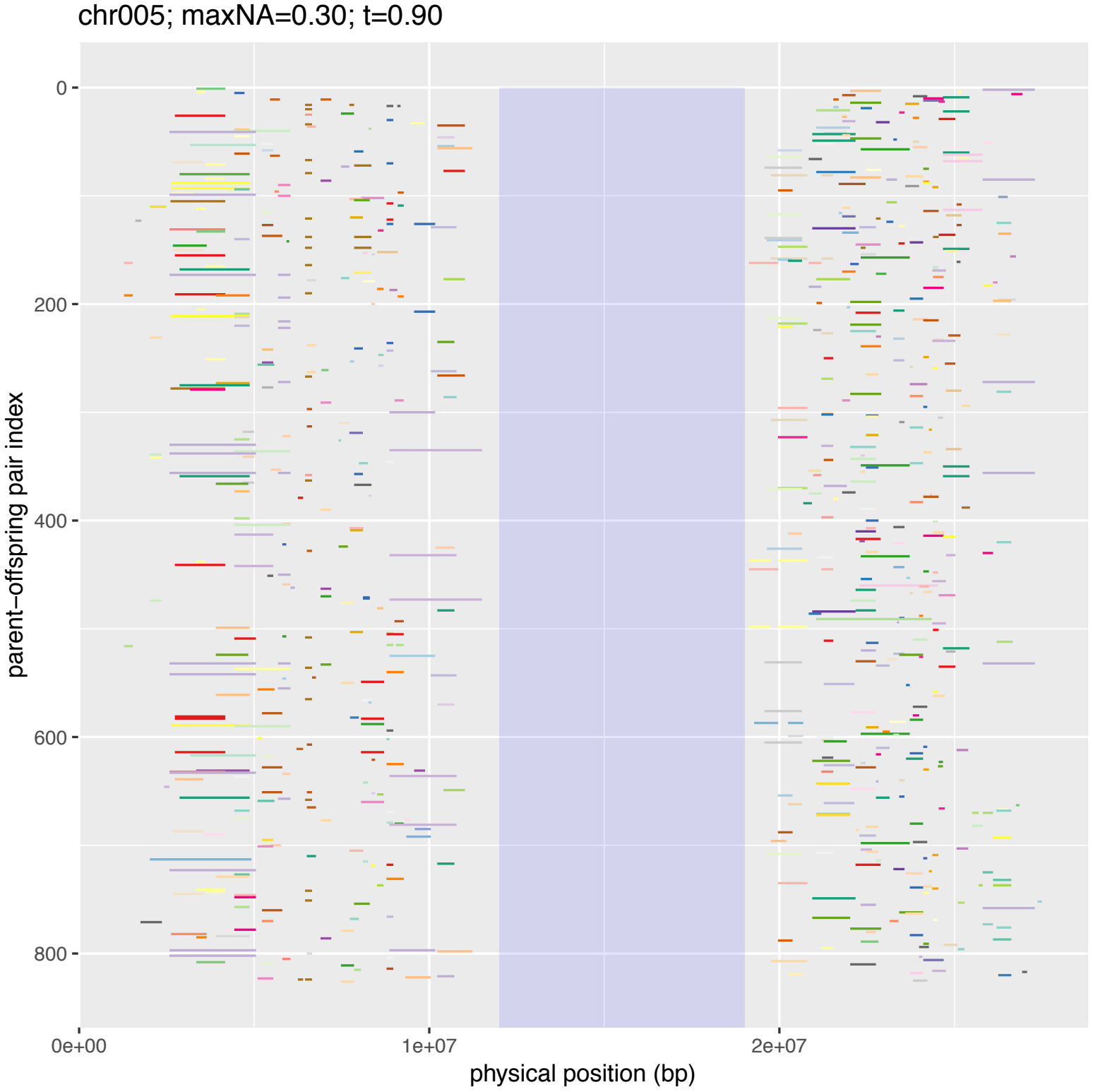

chr018; maxNA=0.20; t=0.90
parent−offspring pair index
0
500
1000
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
physical position (bp)

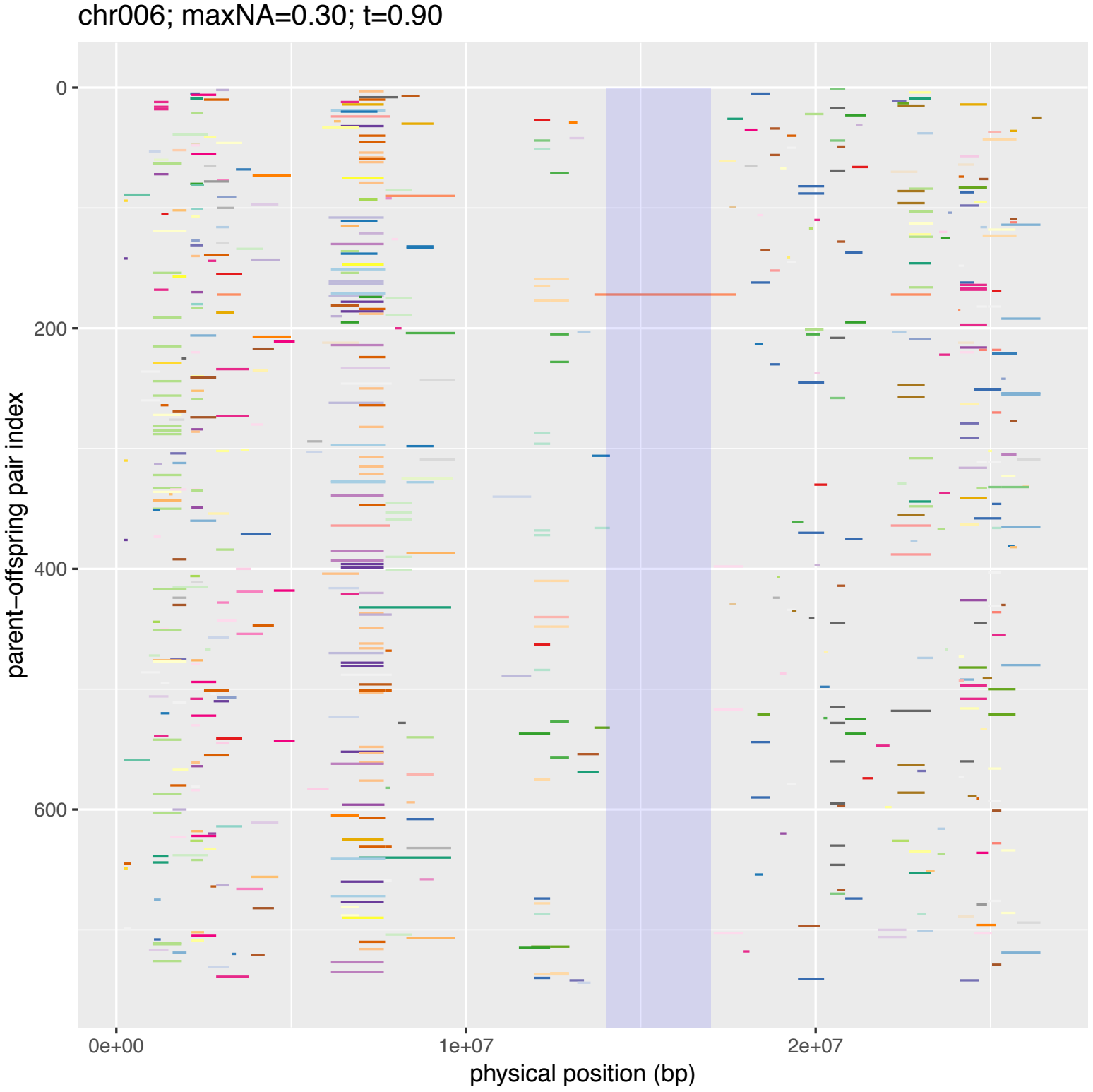

chr001; maxNA=0.30; t=0.90
parent−offspring pair index
0
300
600
900
0e+00
1e+07
2e+07
3e+07
physical position (bp)

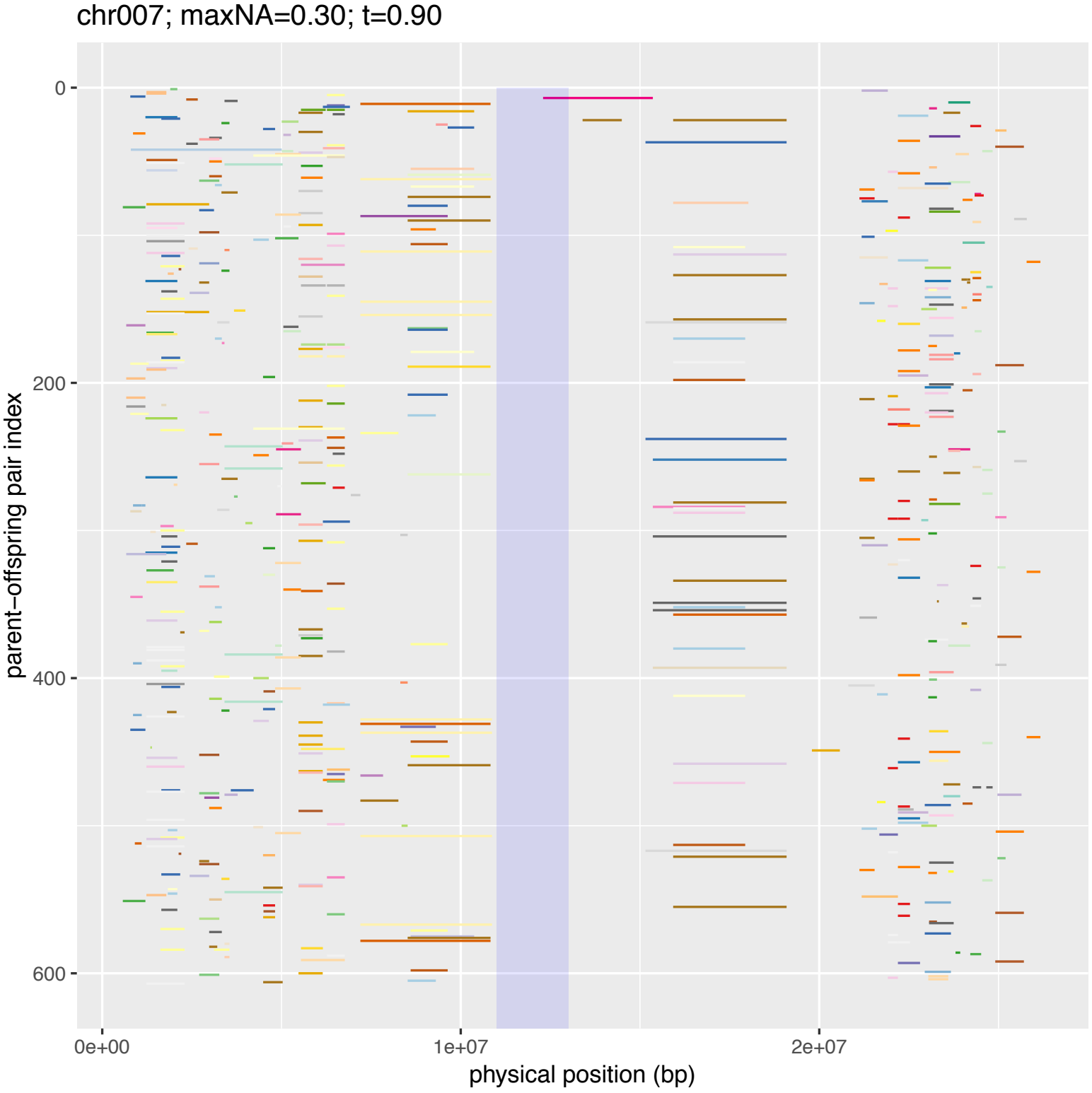

chr002; maxNA=0.30; t=0.90
parent−offspring pair index
0
250
500
750
0e+00
1e+07
2e+07
3e+07
physical position (bp)

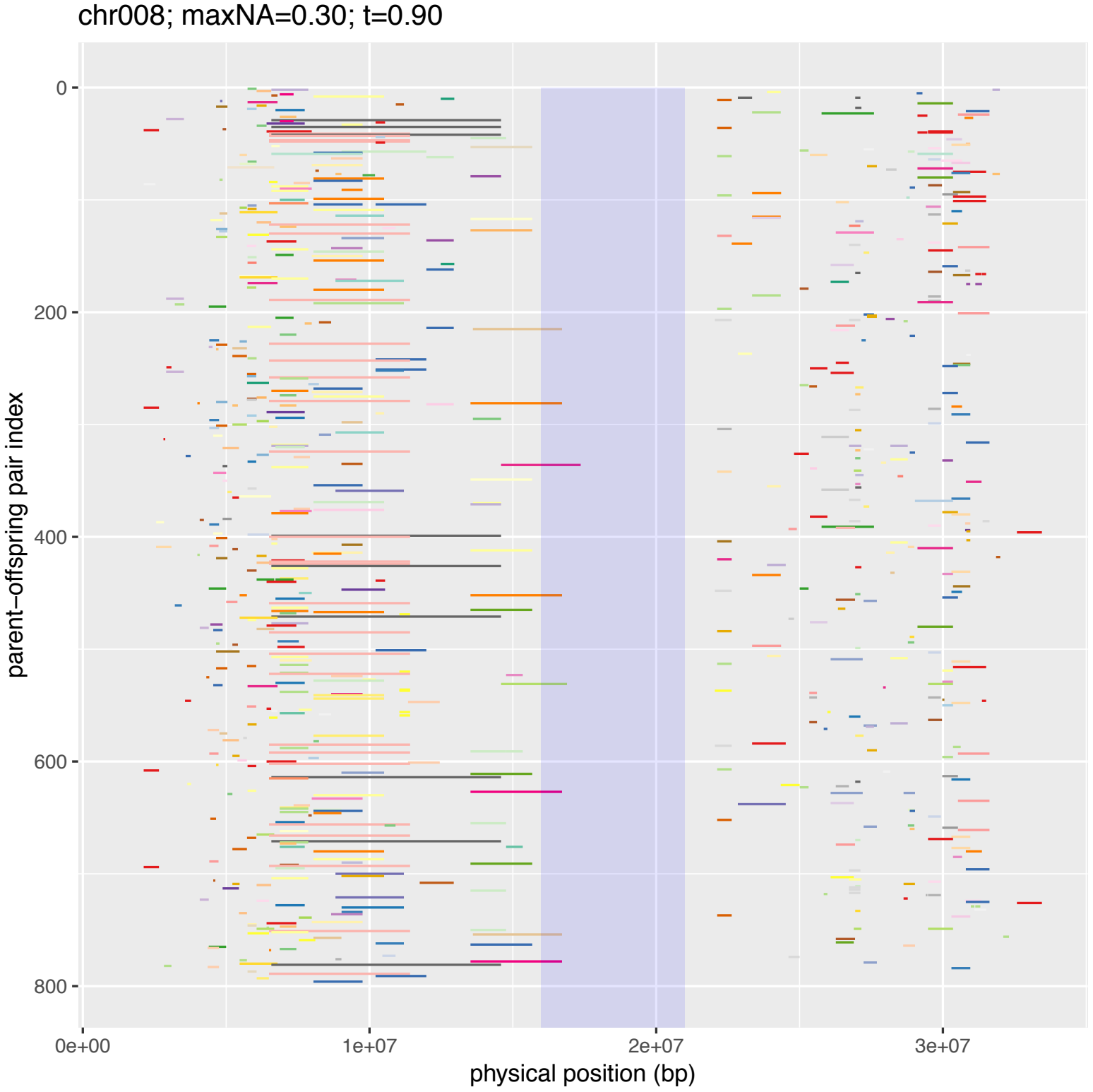

chr003; maxNA=0.30; t=0.90
parent−offspring pair index
0
500
1000
0e+00
1e+07
2e+07
physical position (bp)

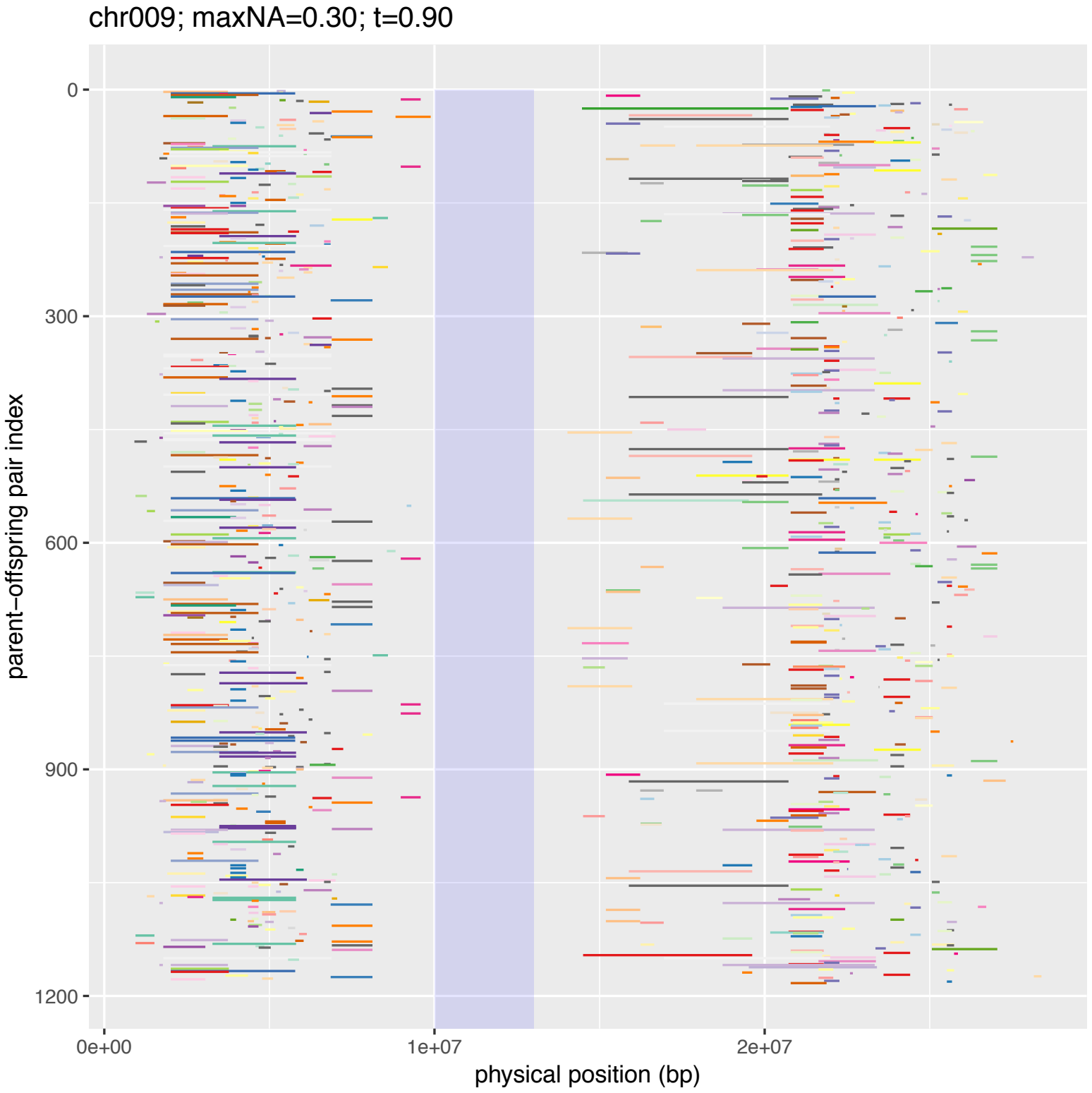

chr004; maxNA=0.30; t=0.90
0
200
400
600
parent−offspring pair index
0e+00
1e+07
2e+07
physical position (bp)

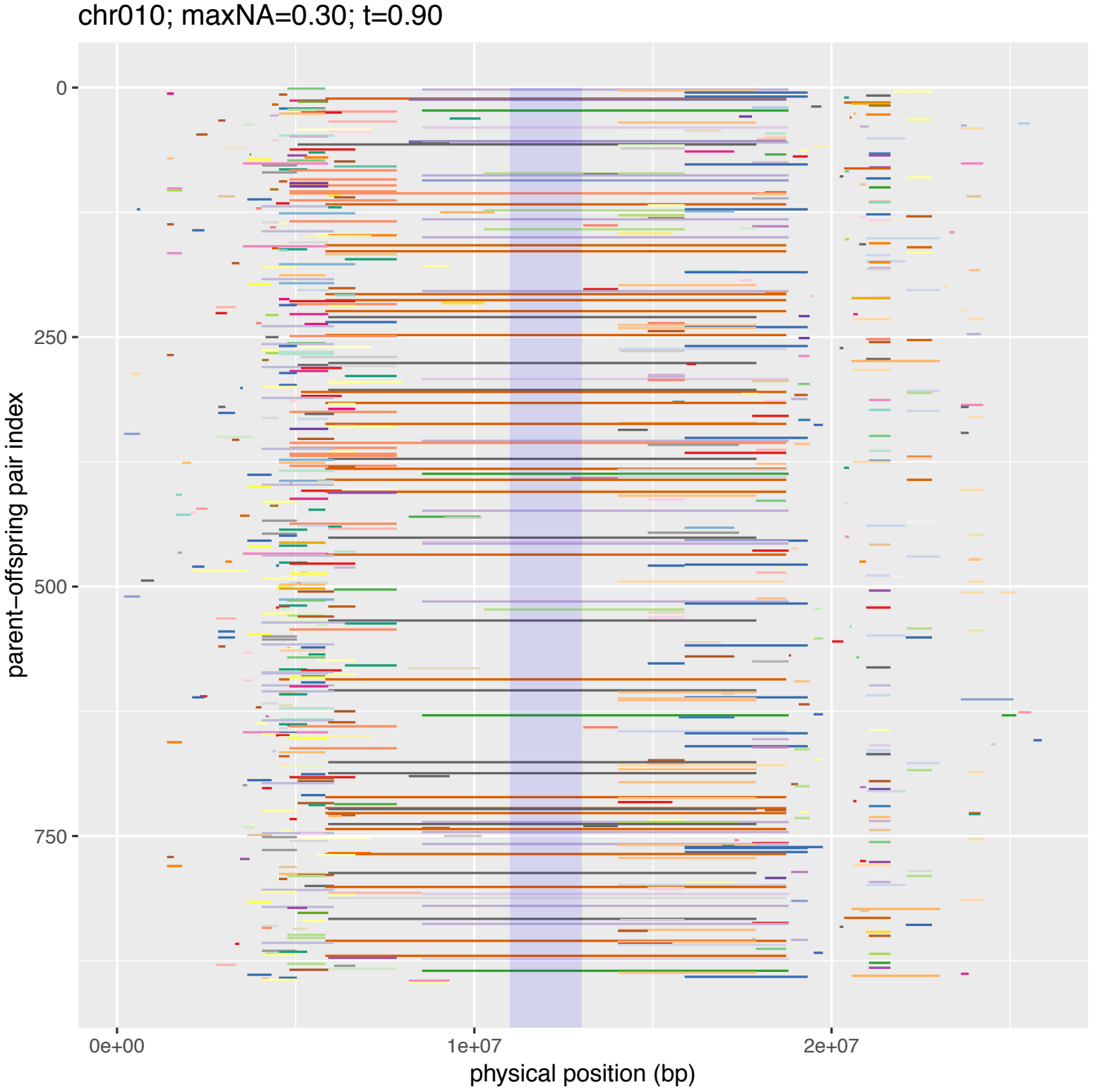

chr005; maxNA=0.30; t=0.90
parent−offspring pair index
0
200
400
600
800
0e+00
1e+07
2e+07
physical position (bp)

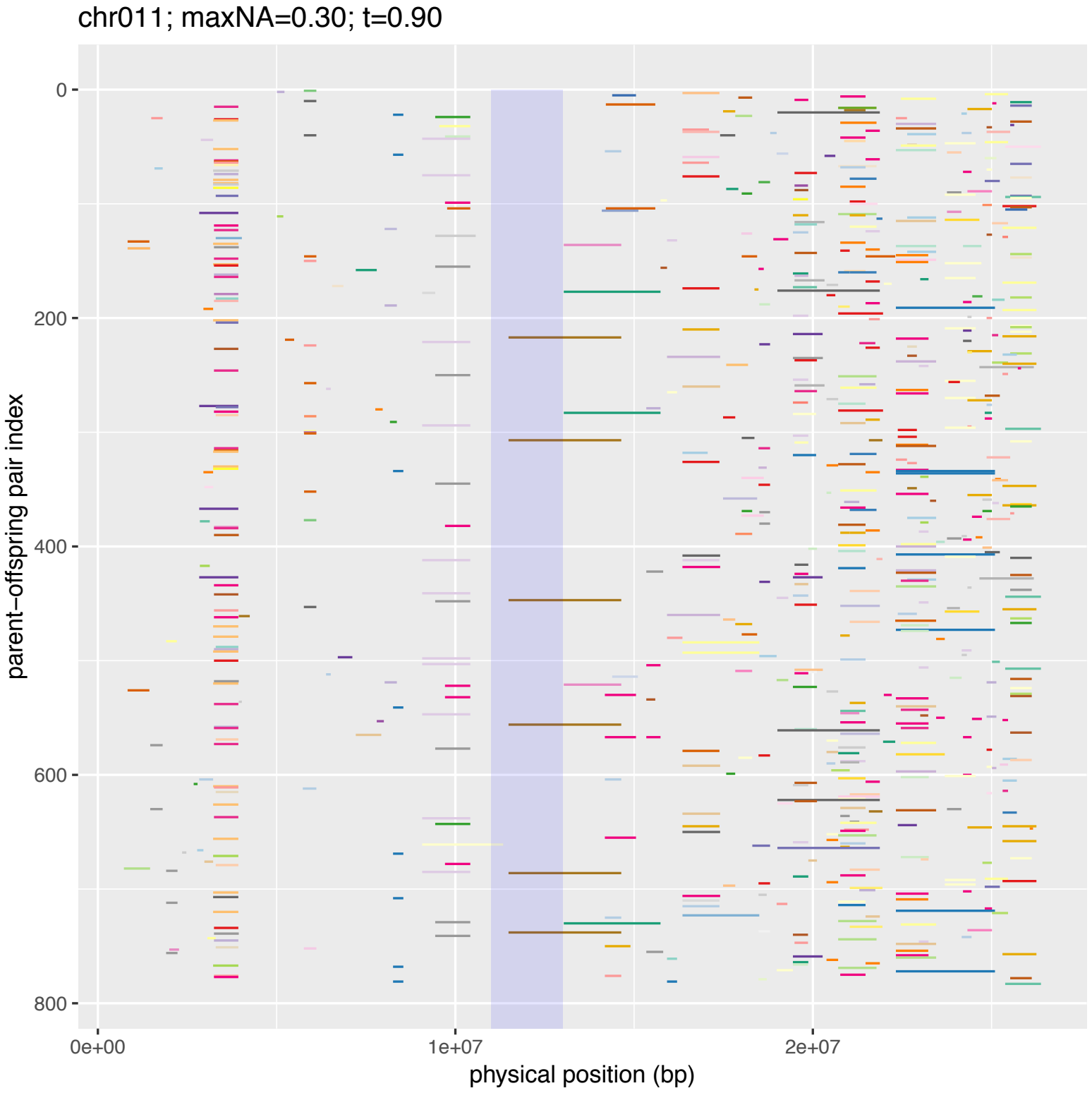

chr006; maxNA=0.30; t=0.90
0
200
400
600
parent−offspring pair index
0e+00
1e+07
2e+07
physical position (bp)

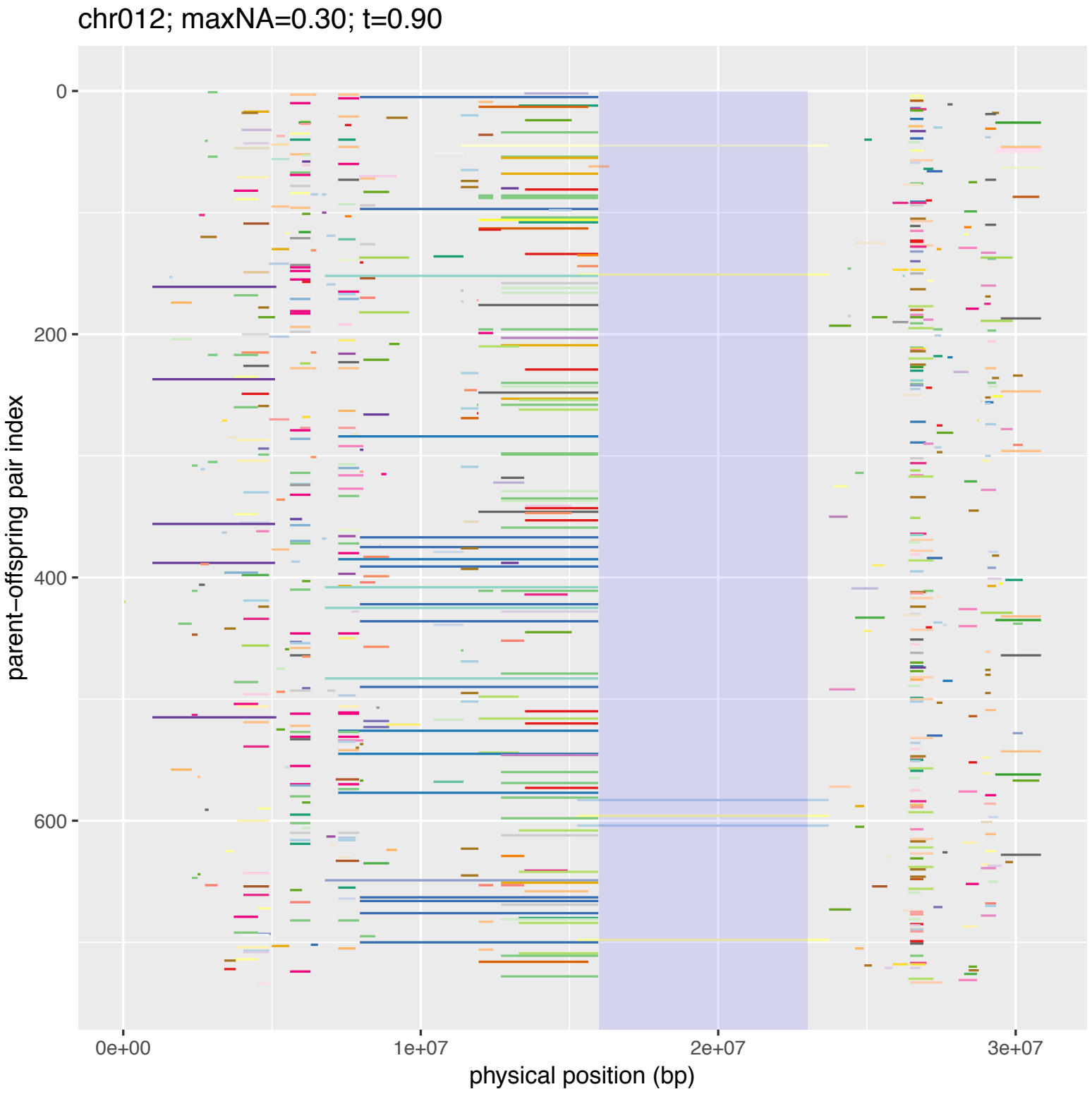

chr007; maxNA=0.30; t=0.90
0
200
400
600
parent−offspring pair index
0e+00
1e+07
2e+07
physical position (bp)

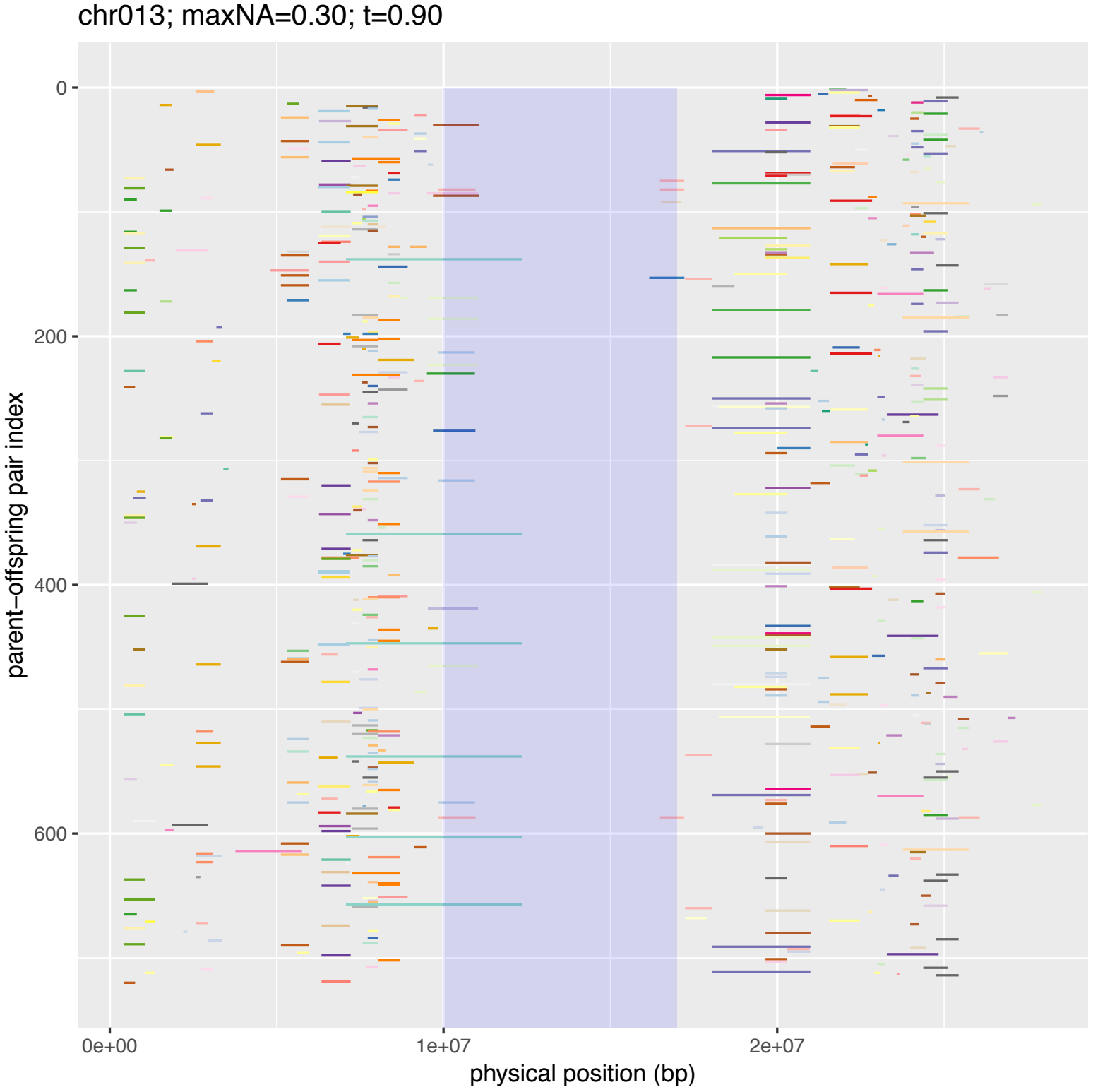

chr008; maxNA=0.30; t=0.90
parent−offspring pair index
0
200
400
600
800
0e+00
1e+07
2e+07
3e+07
physical position (bp)

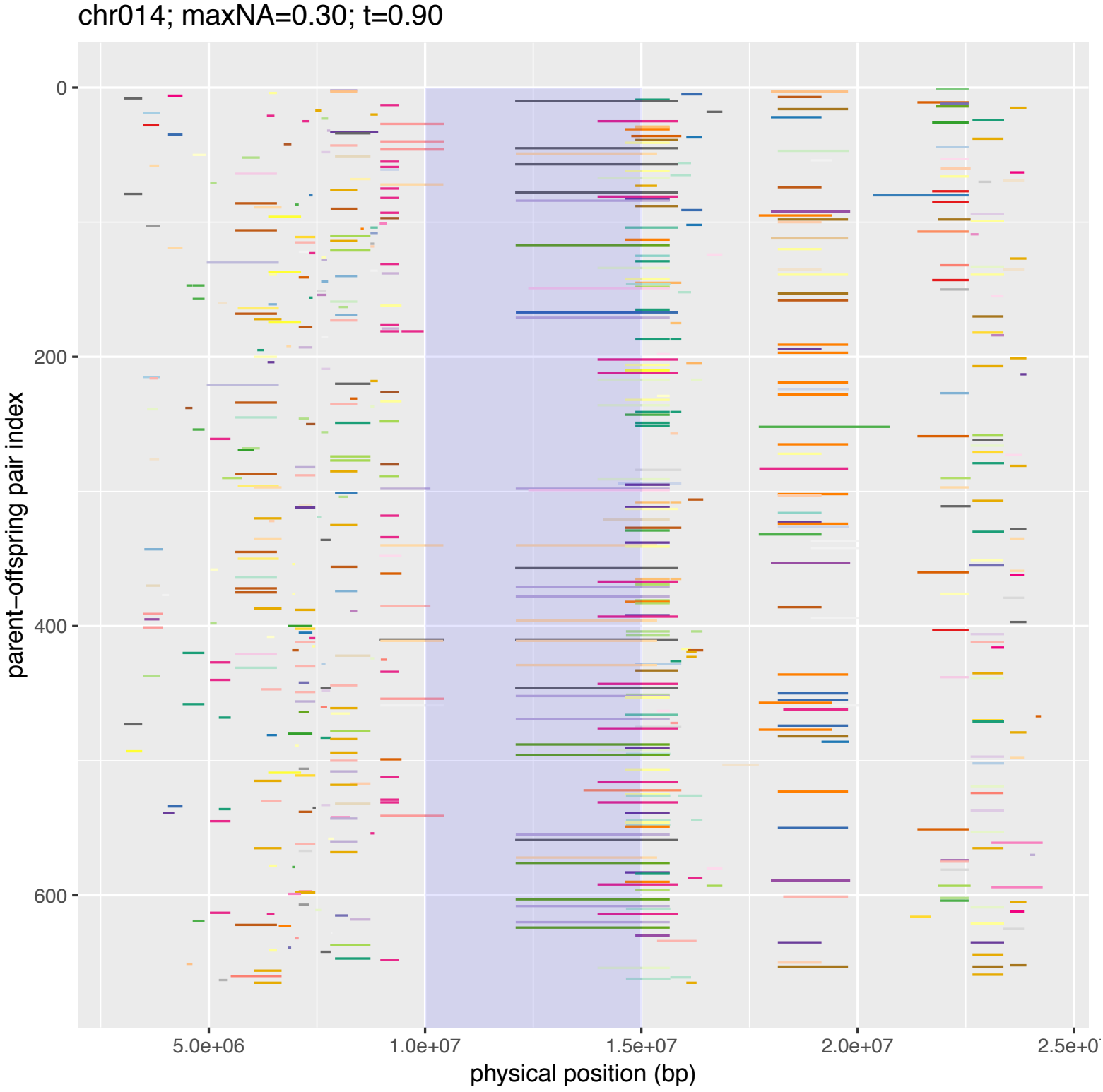

chr009; maxNA=0.30; t=0.90
parent−offspring pair index
0
300
600
900
1200
0e+00
1e+07
2e+07
physical position (bp)

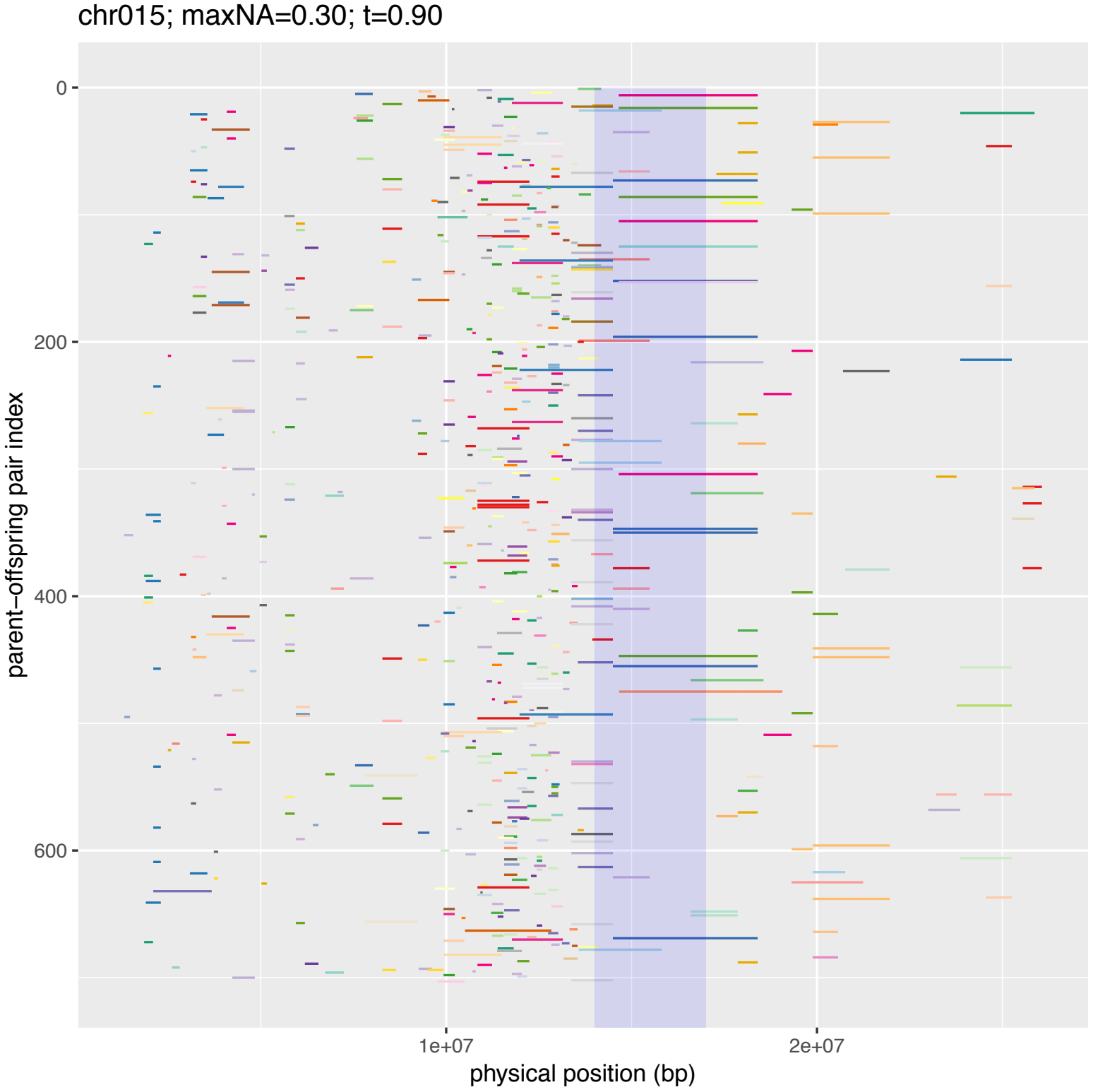

chr010; maxNA=0.30; t=0.90
0
250
500
750
parent−offspring pair index
0e+00
1e+07
2e+07
physical position (bp)

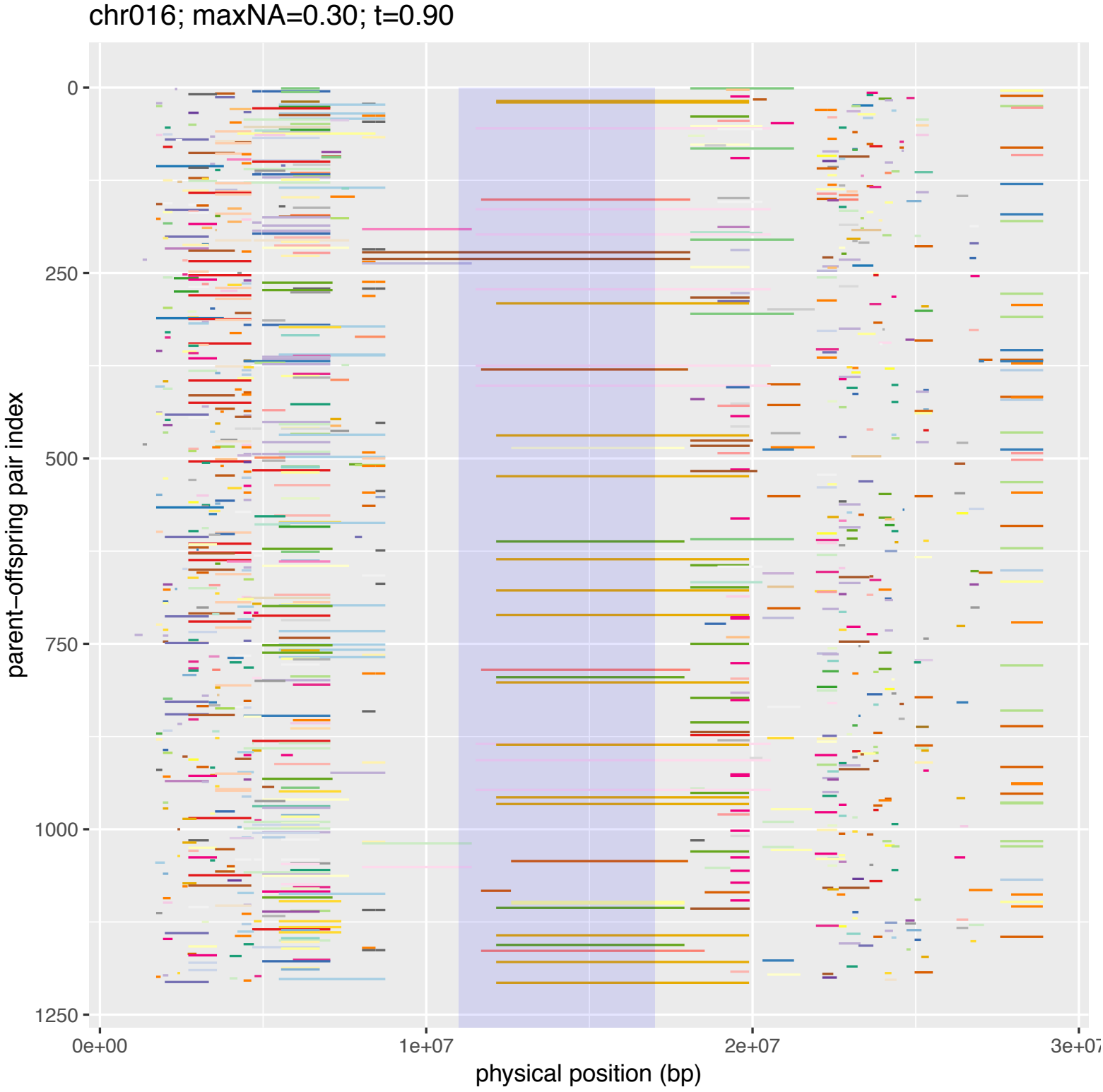

chr011; maxNA=0.30; t=0.90
parent−offspring pair index
0
200
400
600
800
0e+00
1e+07
2e+07
physical position (bp)

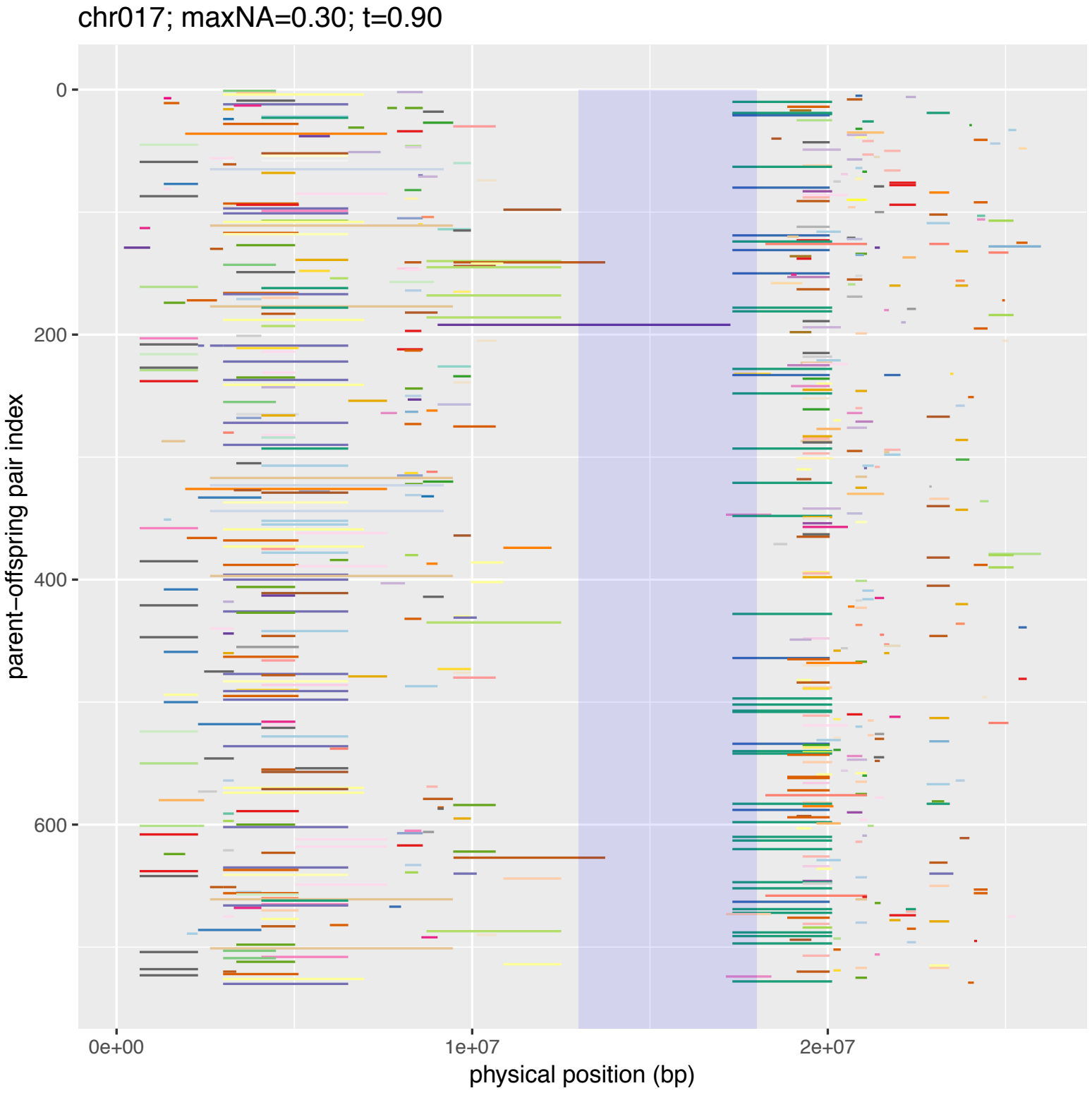

chr012; maxNA=0.30; t=0.90
parent−offspring pair index
0
200
400
600
0e+00
1e+07
2e+07
3e+07
physical position (bp)

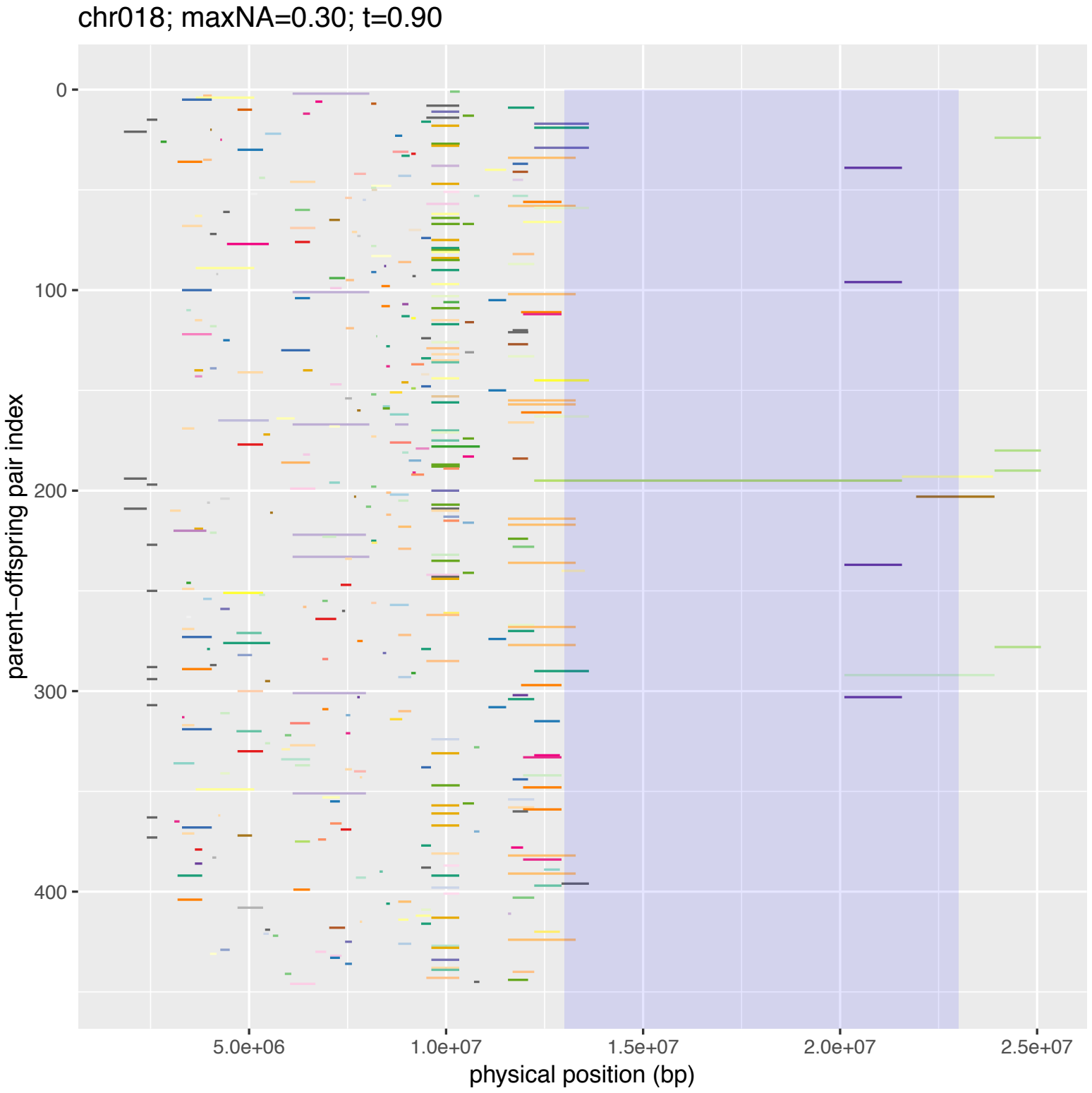

chr013; maxNA=0.30; t=0.90
0
200
400
600
parent−offspring pair index
0e+00
1e+07
2e+07
physical position (bp)


chr014; maxNA=0.30; t=0.90
parent−offspring pair index
0
200
400
600
5.0e+06
1.0e+07
1.5e+07
2.0e+07
physical position (bp)


chr015; maxNA=0.30; t=0.90
0
200
400
600
parent−offspring pair index
1e+07
2e+07
physical position (bp)

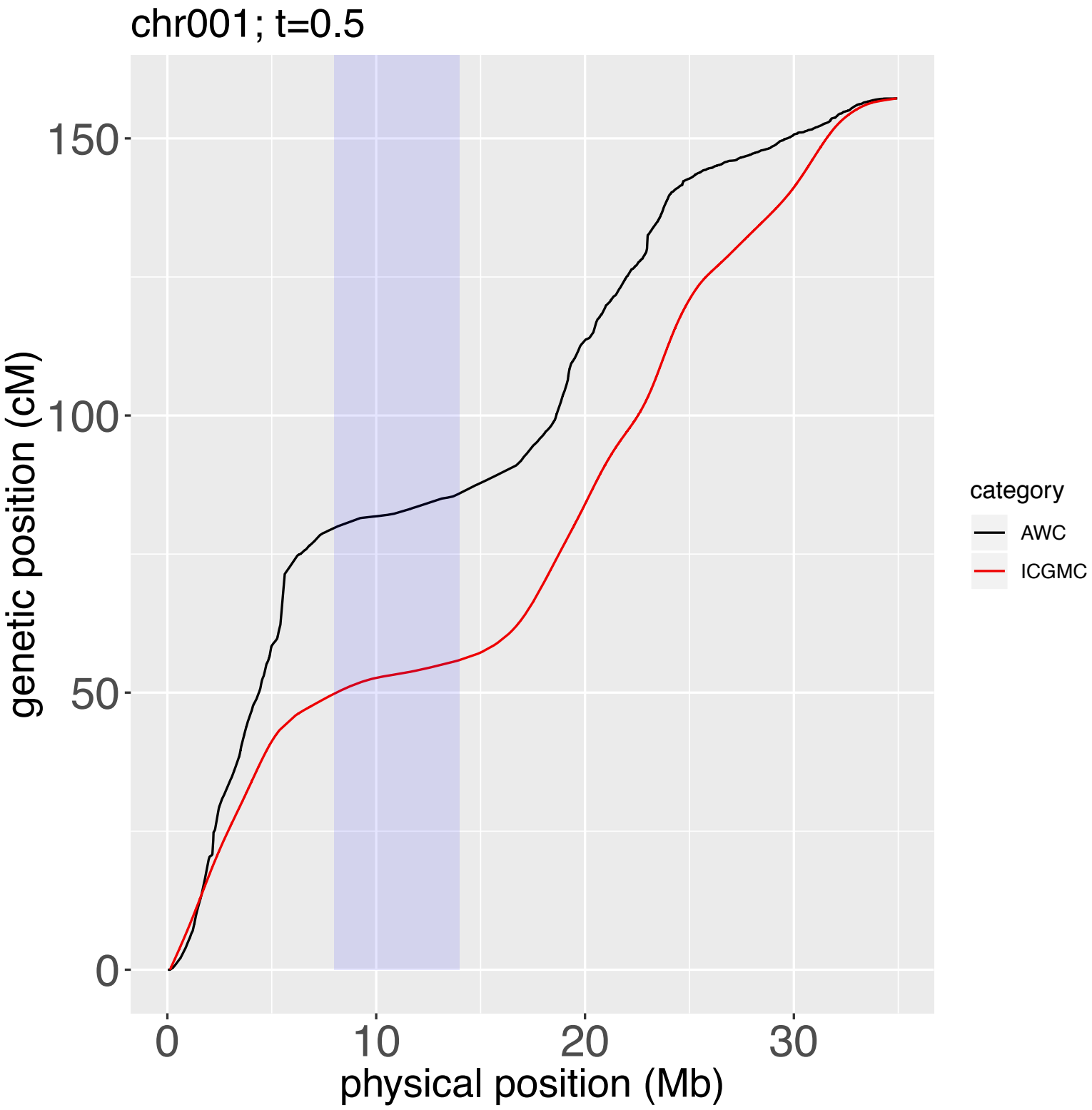

chr016; maxNA=0.30; t=0.90
parent−offspring pair index
0
250
500
750
1000
1250
0e+00
1e+07
2e+07
3e+07
physical position (bp)

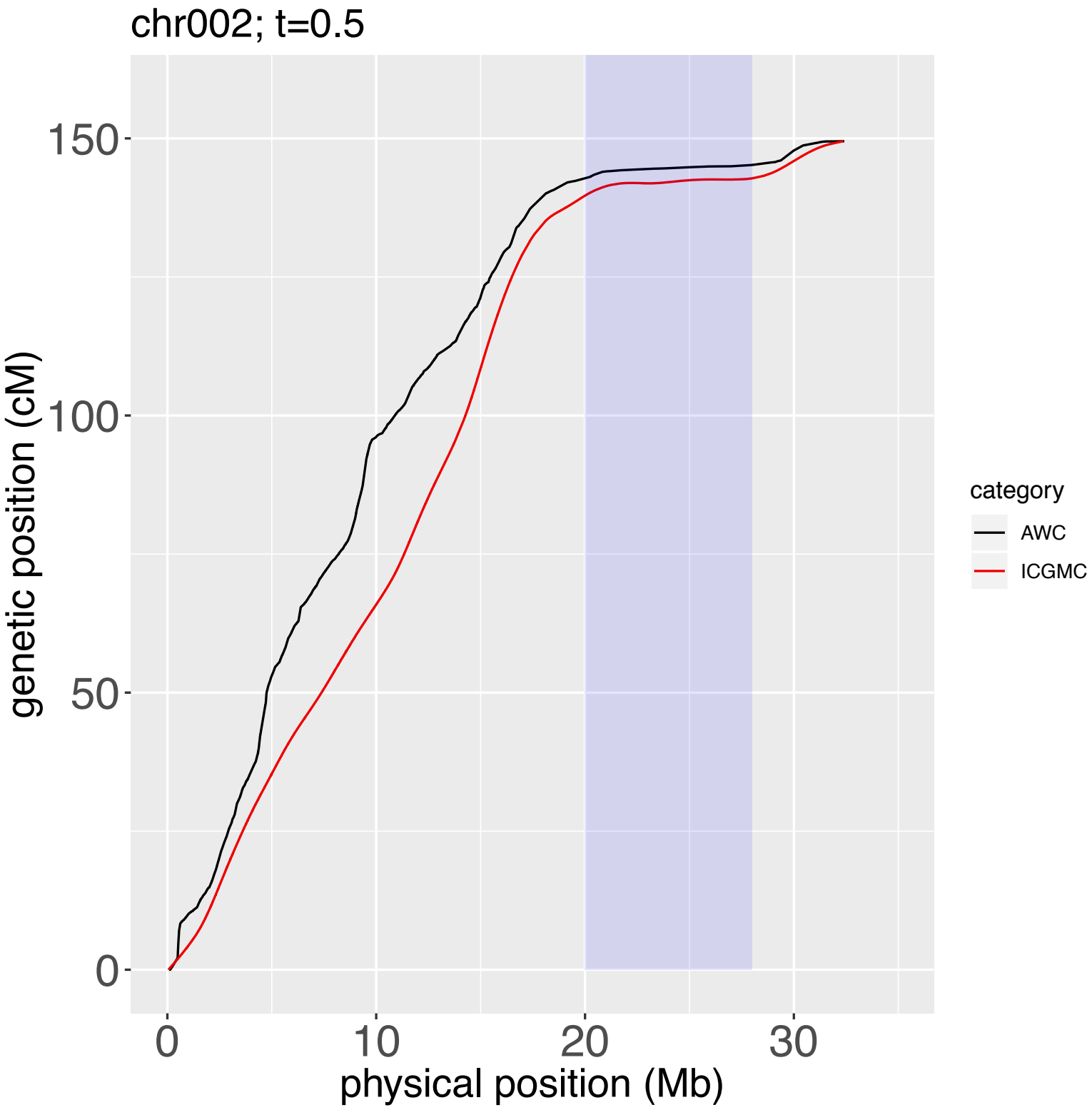

chr017; maxNA=0.30; t=0.90
0
200
400
600
parent−offspring pair index
0e+00
1e+07
2e+07
physical position (bp)

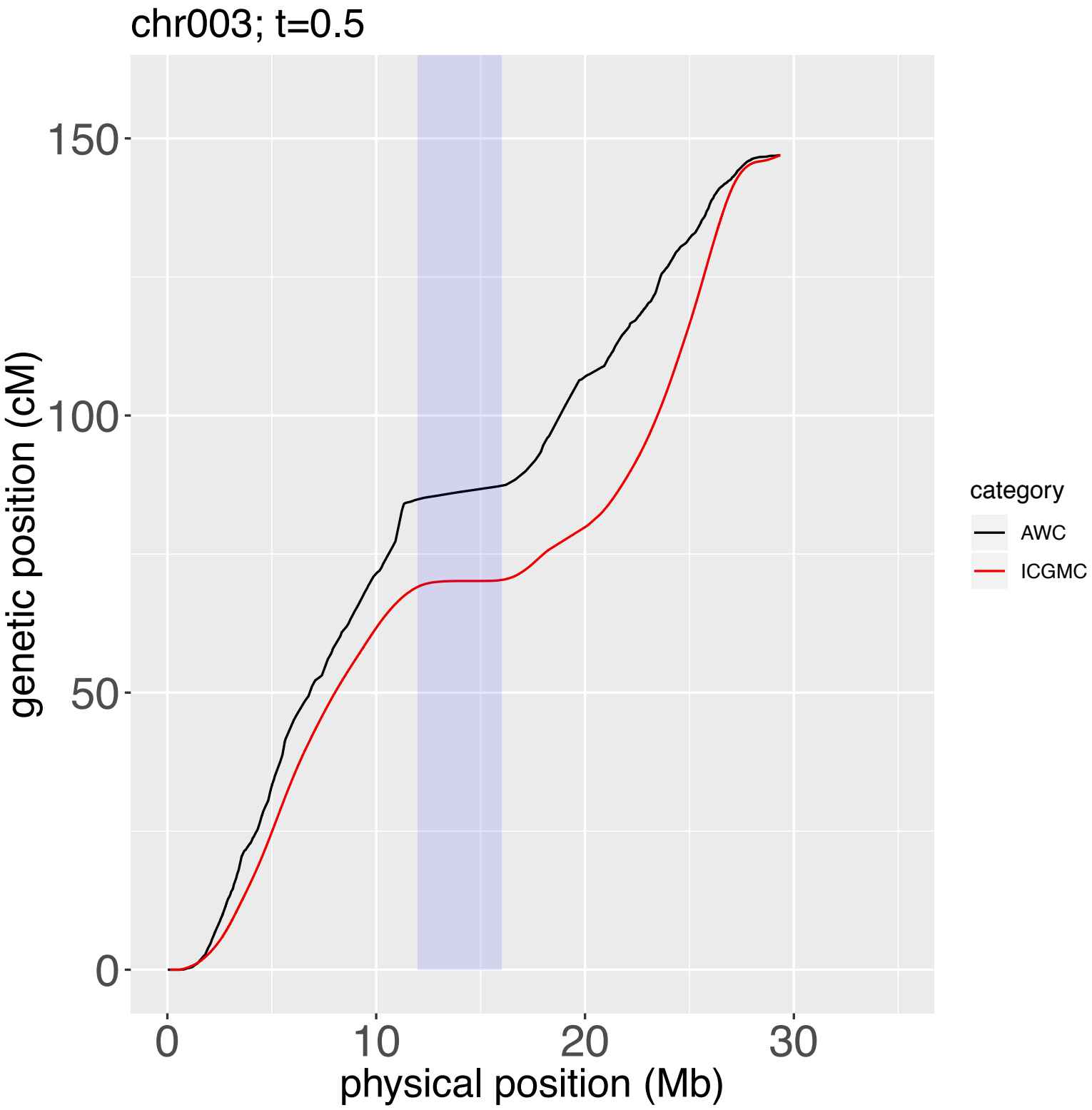

chr018; maxNA=0.30; t=0.90
parent−offspring pair index
0
100
200
300
400
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
physical position (bp)

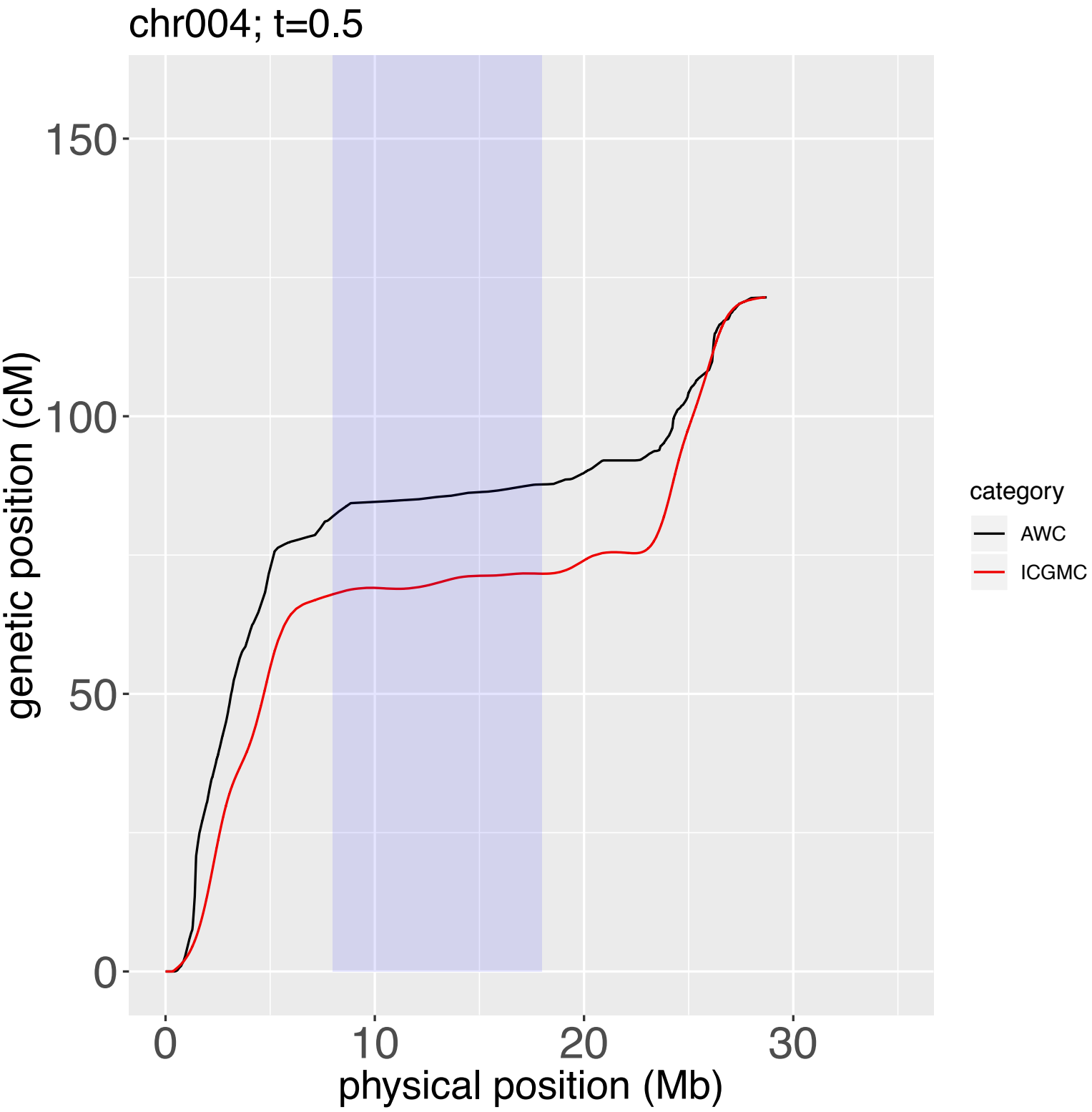
Appendix Figure 3.2

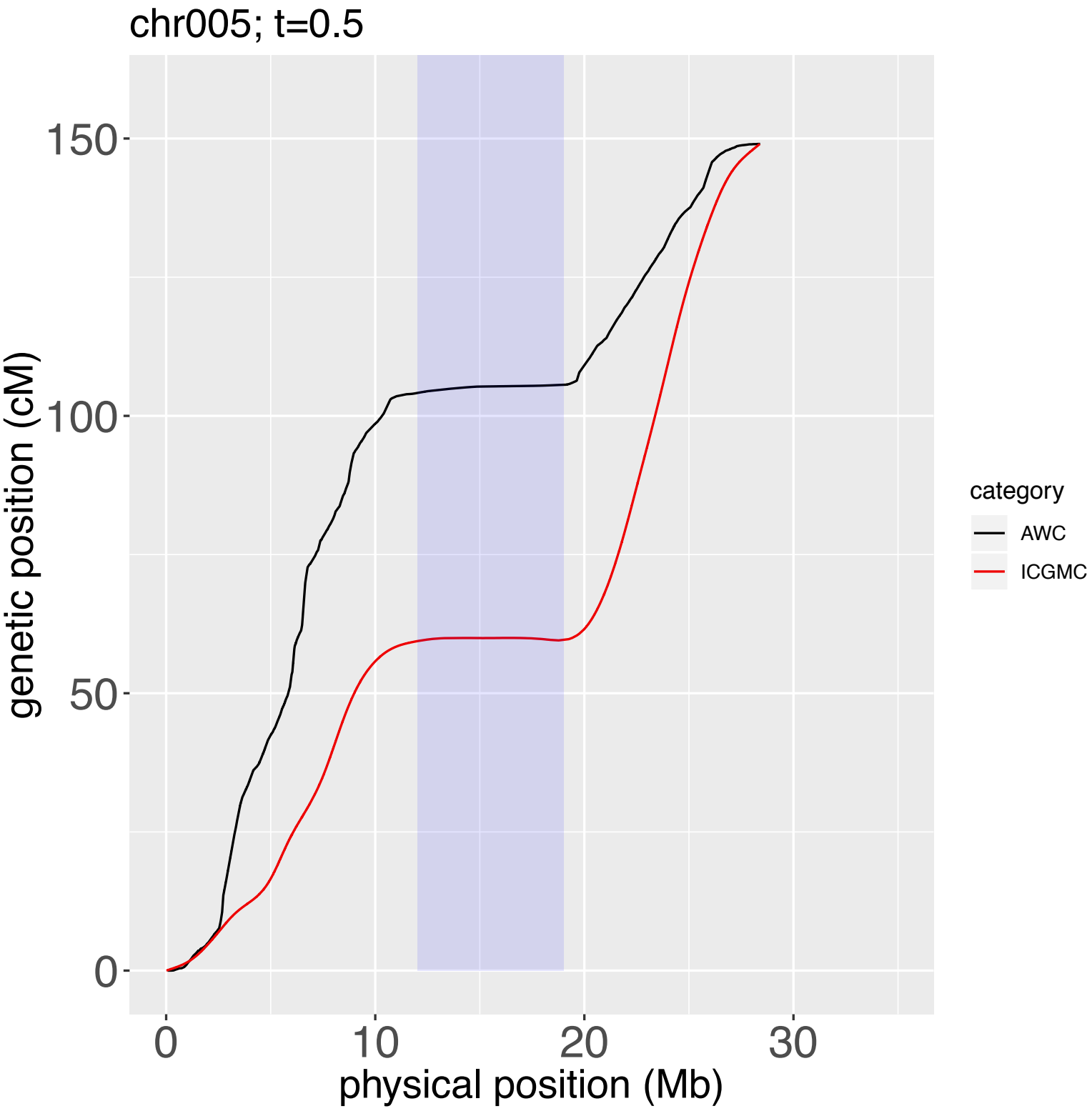
To compare our map to ICGMC's, we plotted the genetic position (cM) of our markers (black) and ICGMC's markers (red) as a function of physical position (Mb). We scaled our map by a factor of three. This figure includes a plot for each chromosome.

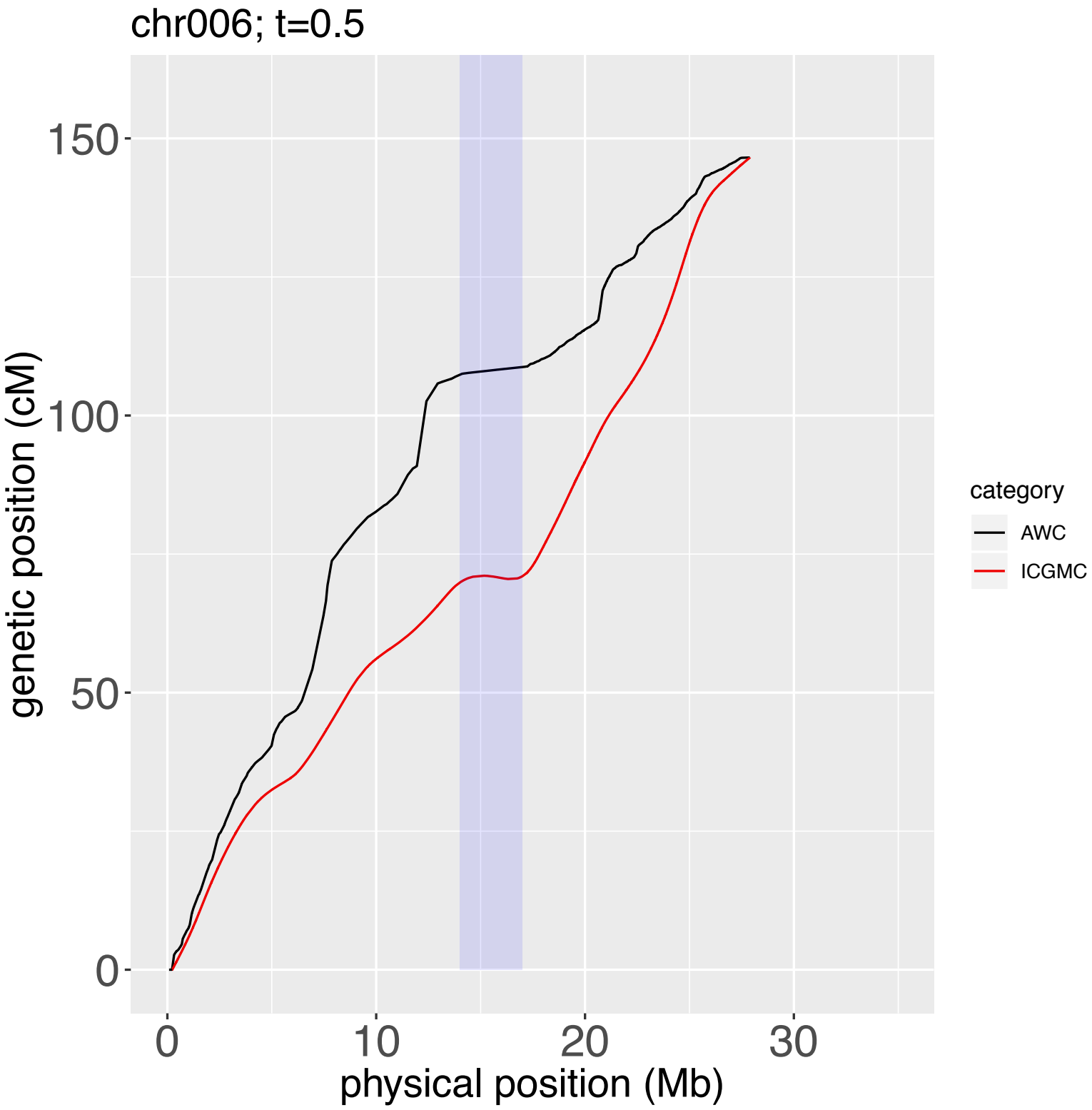

chr001; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

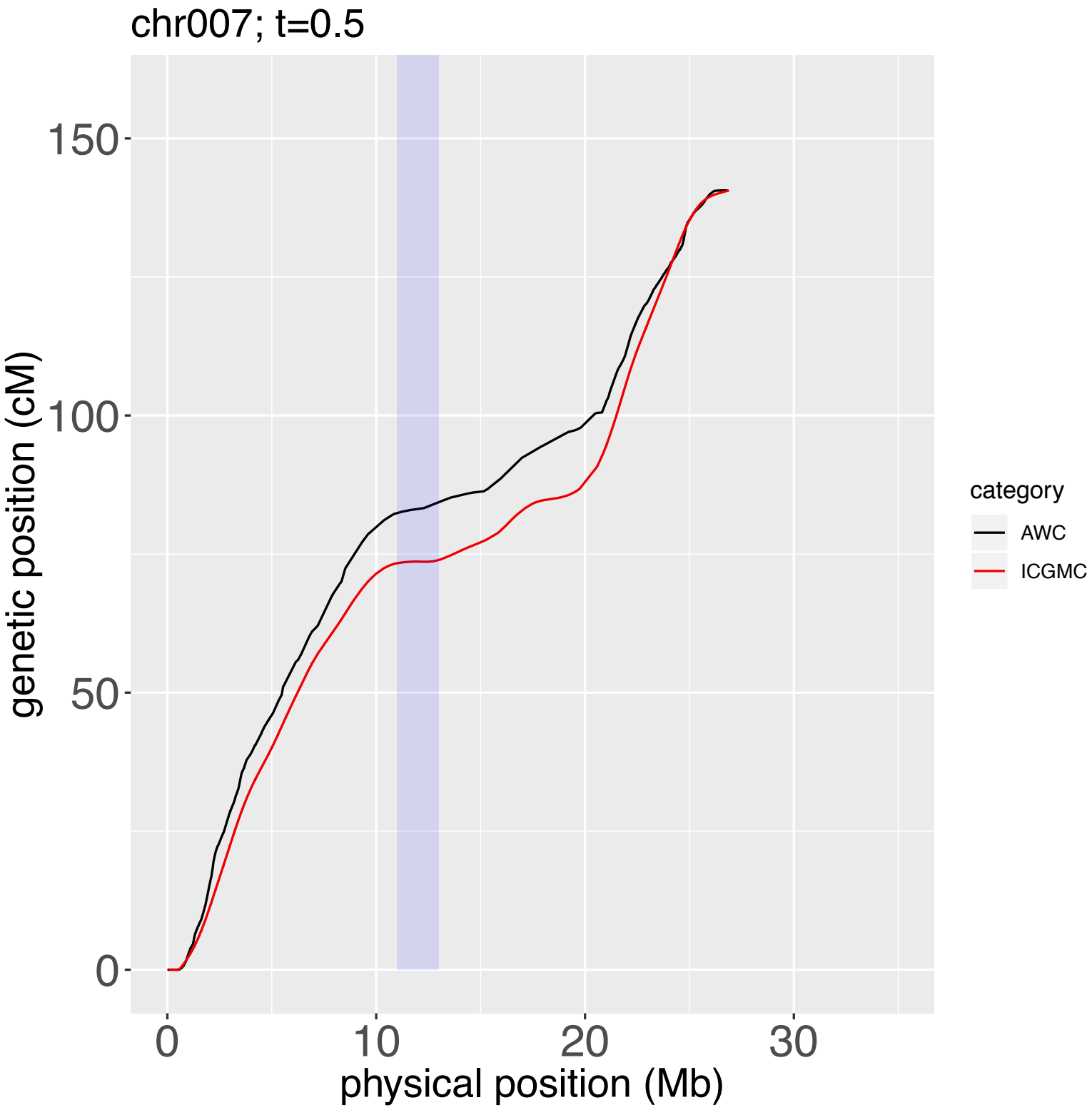

chr002; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

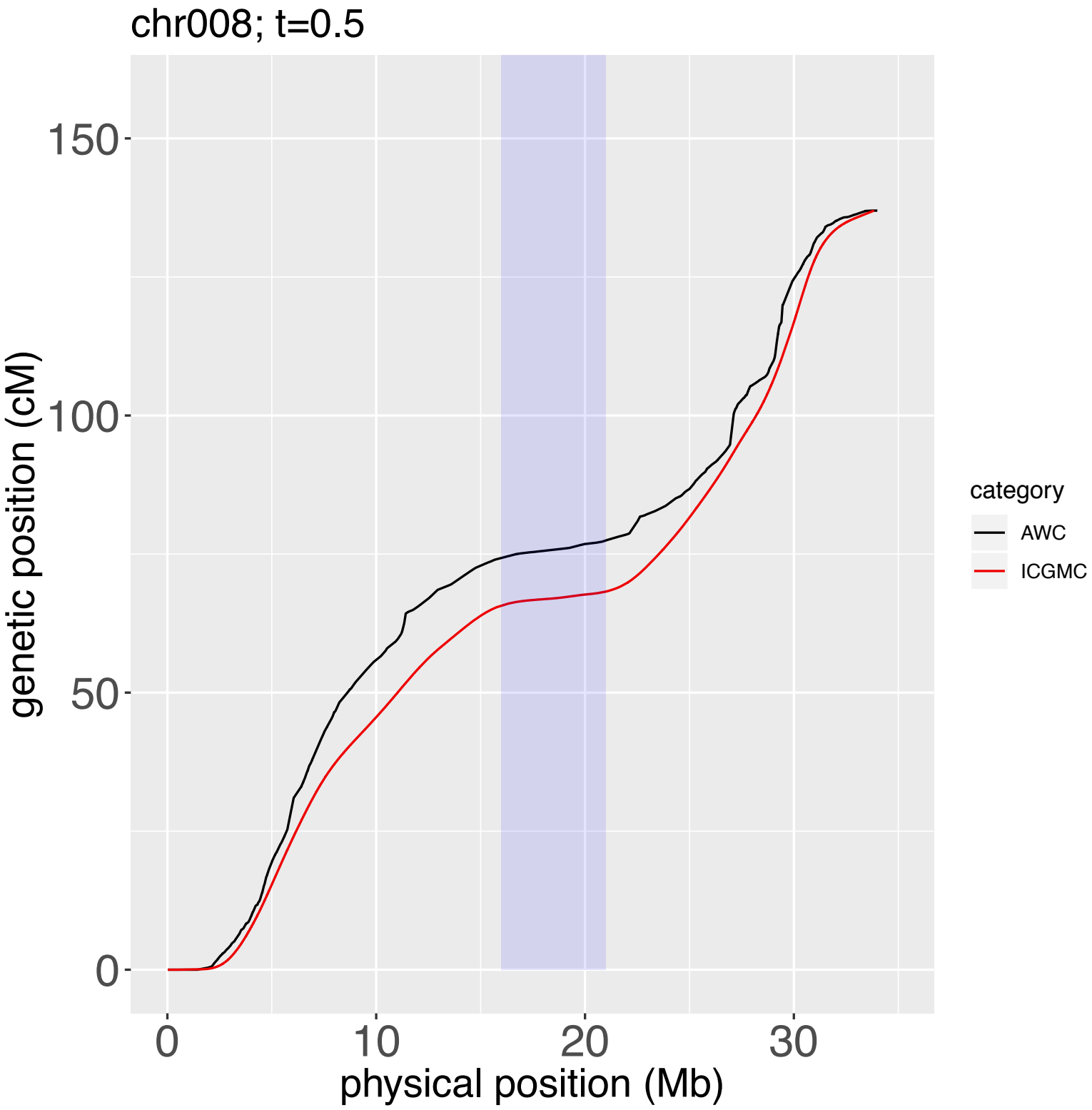

chr003; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

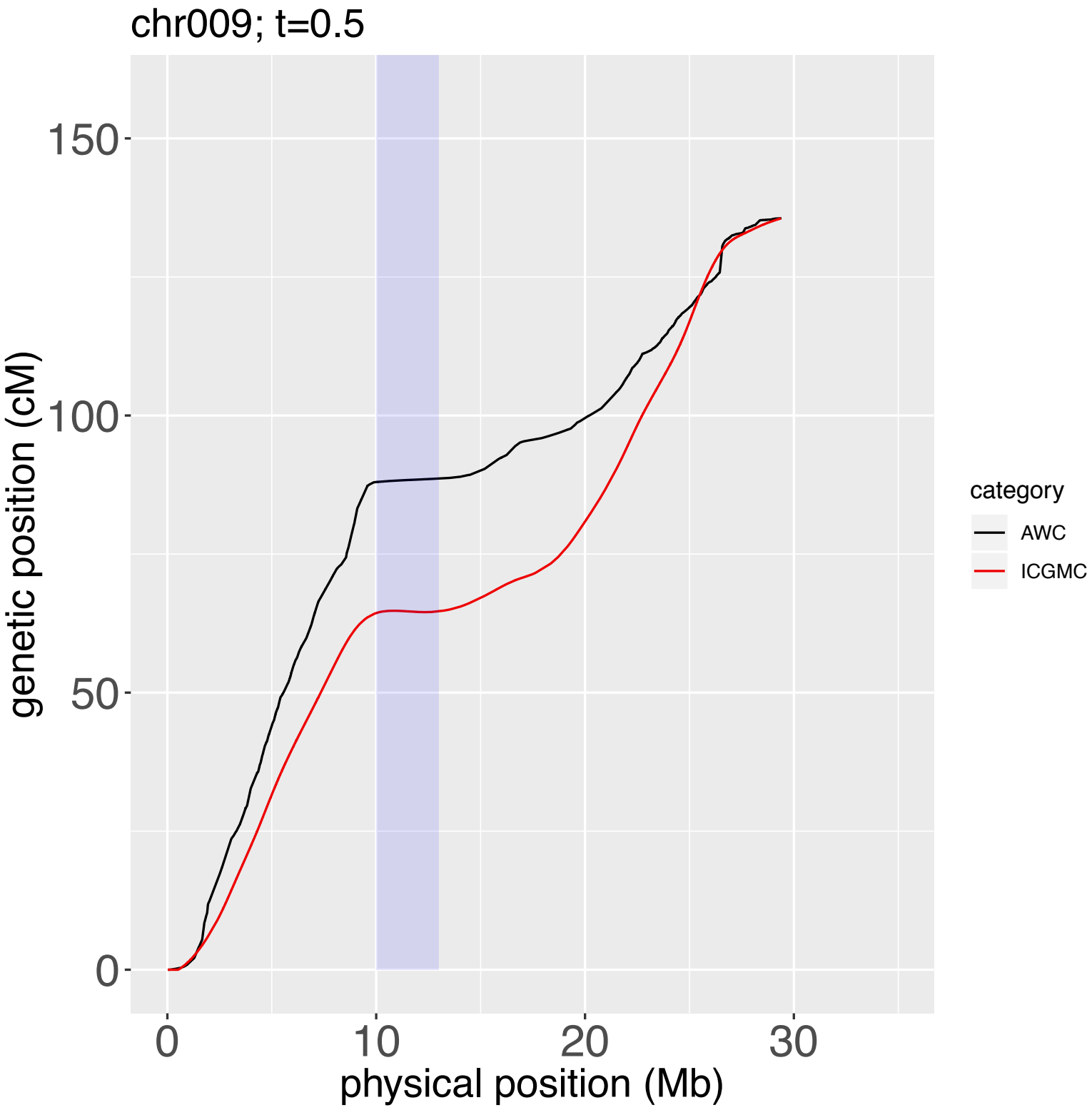

chr004; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

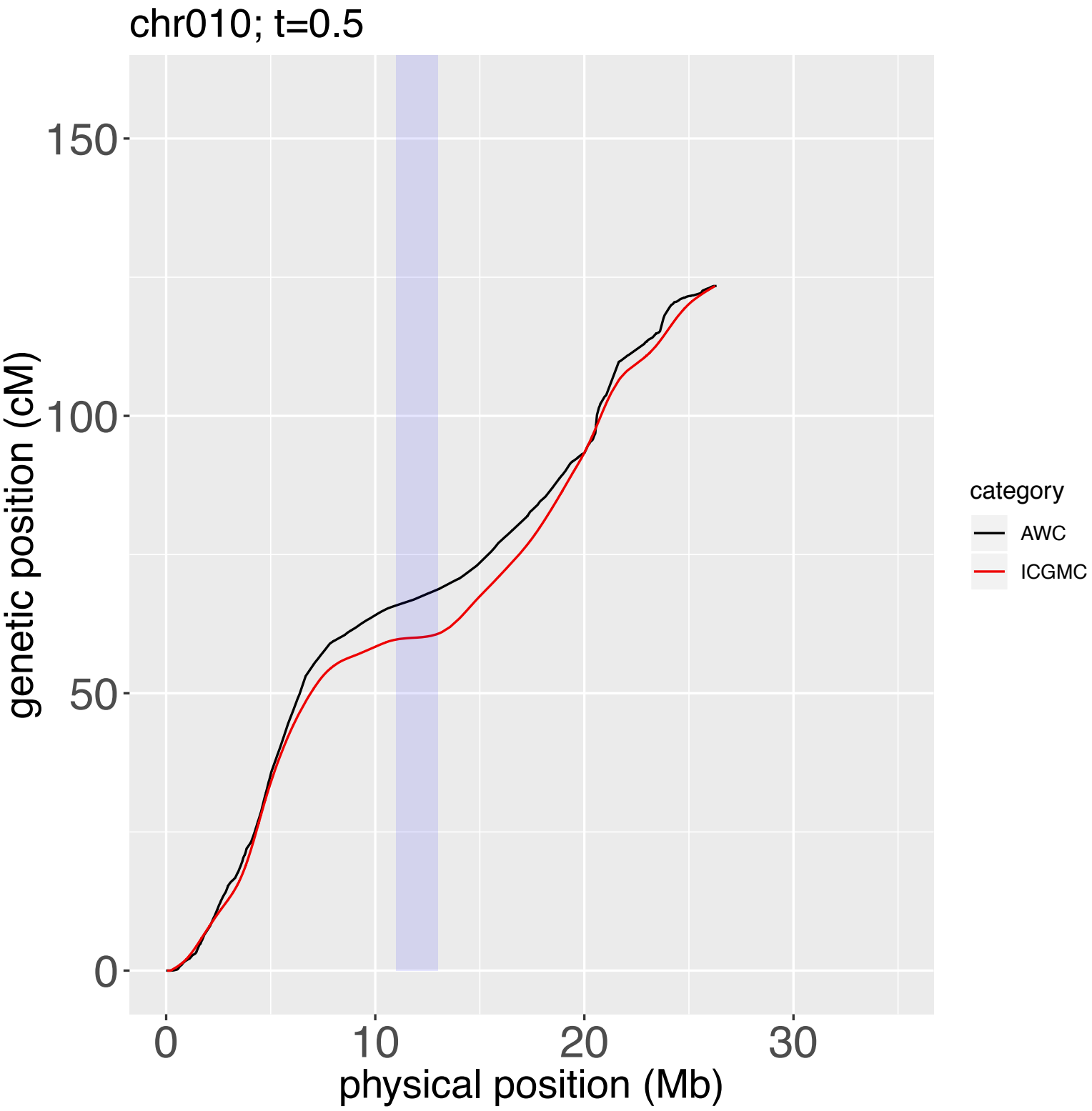

chr005; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

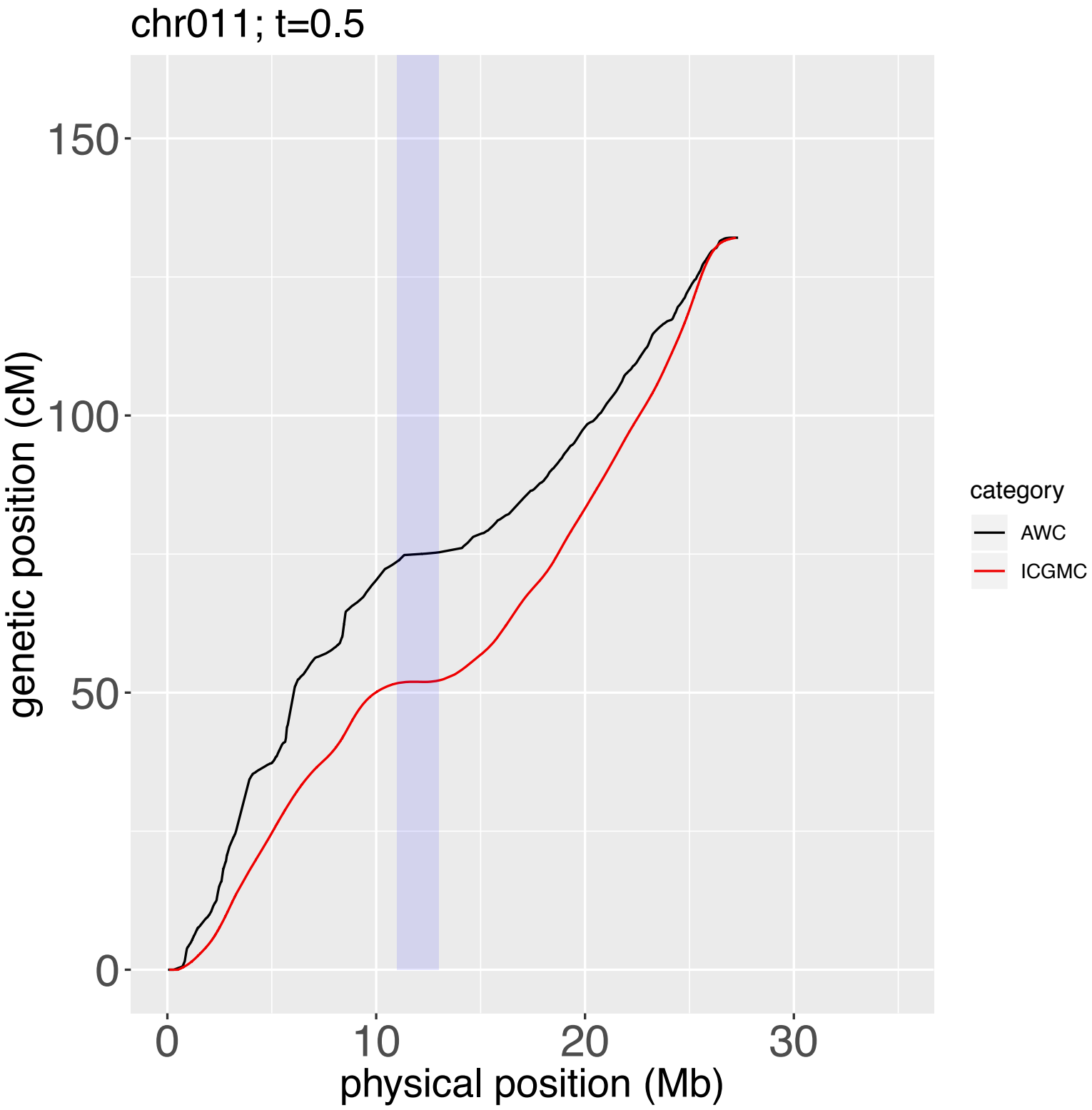

chr006; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

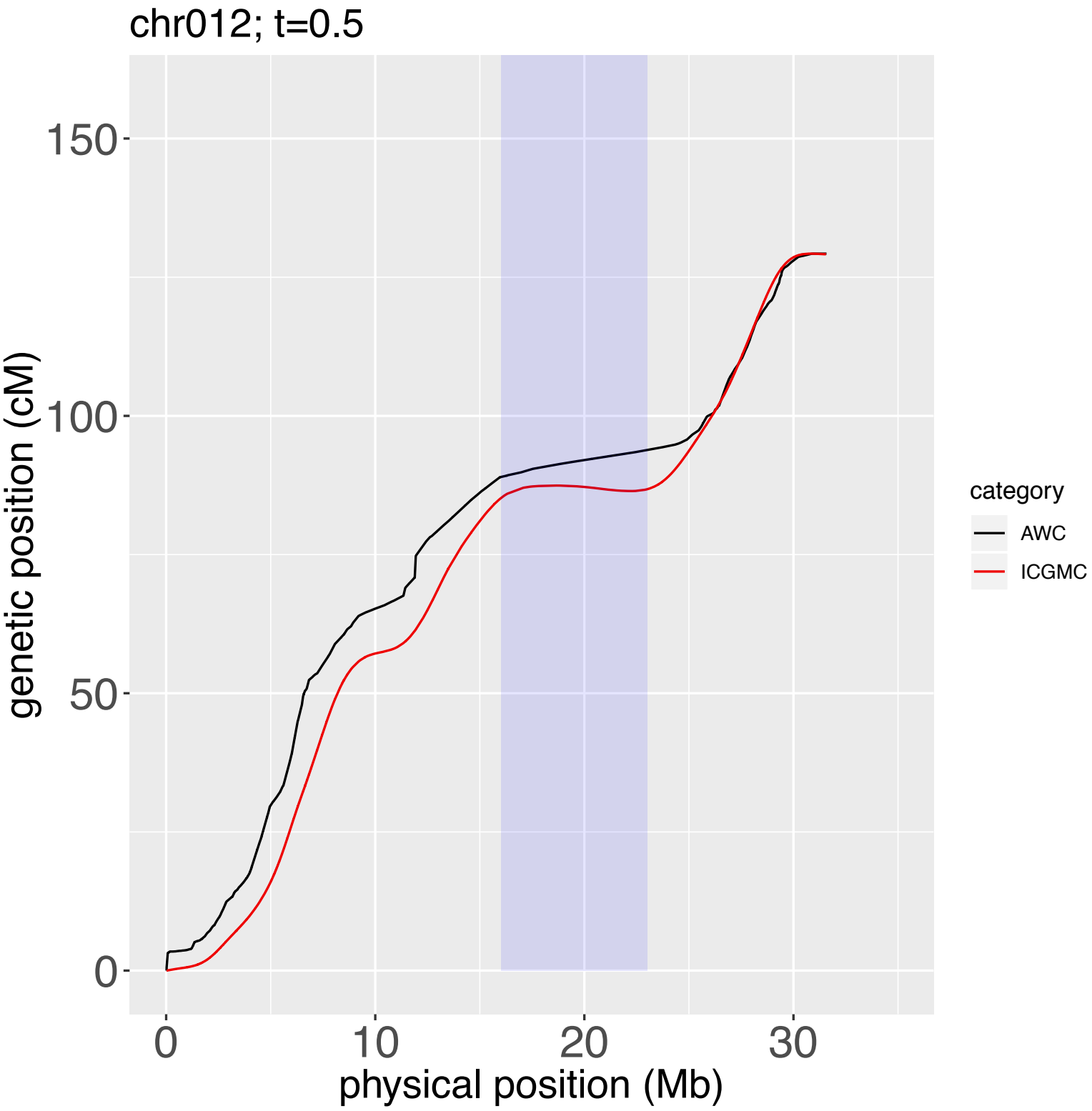

chr007; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

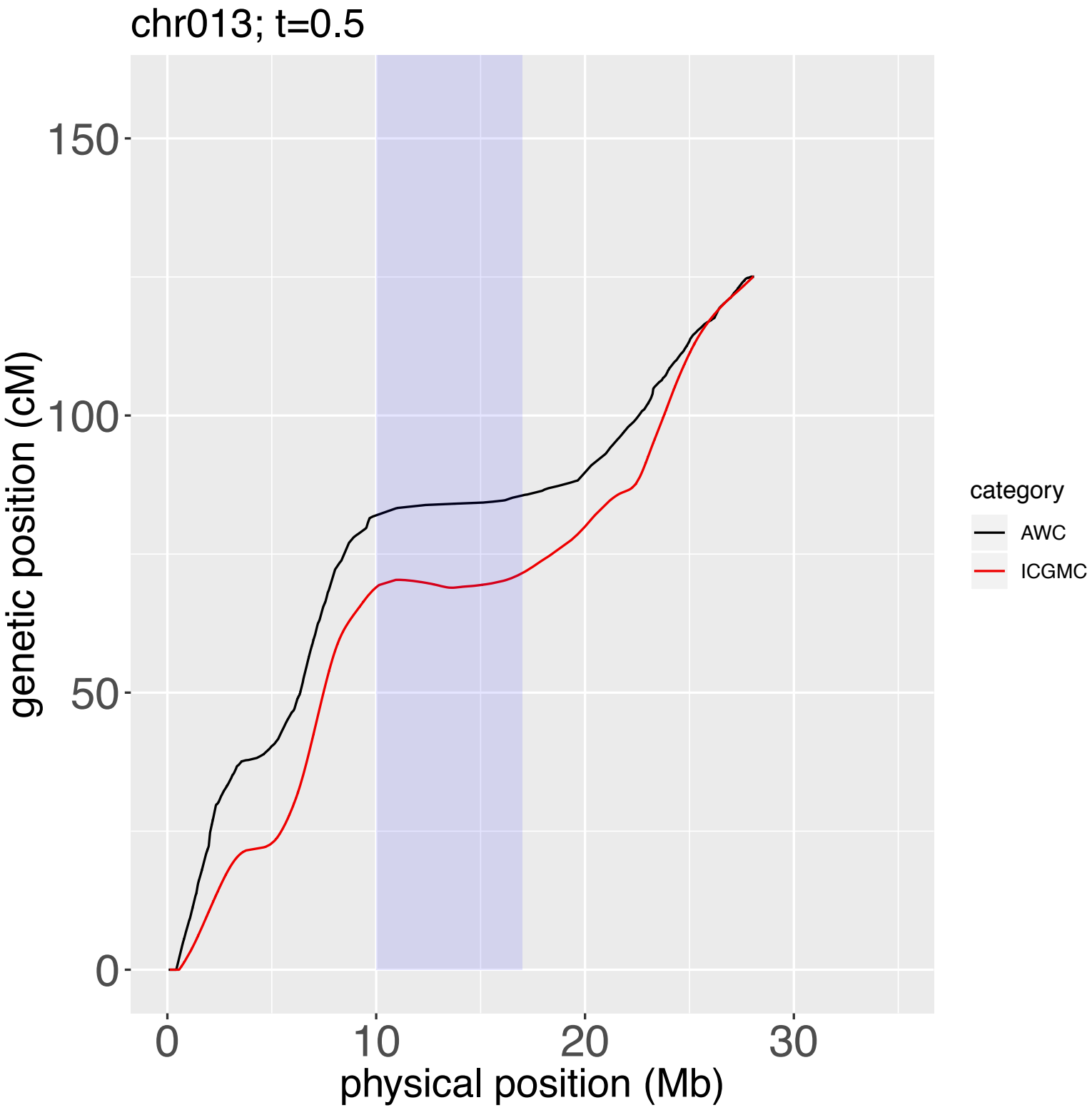

chr008; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

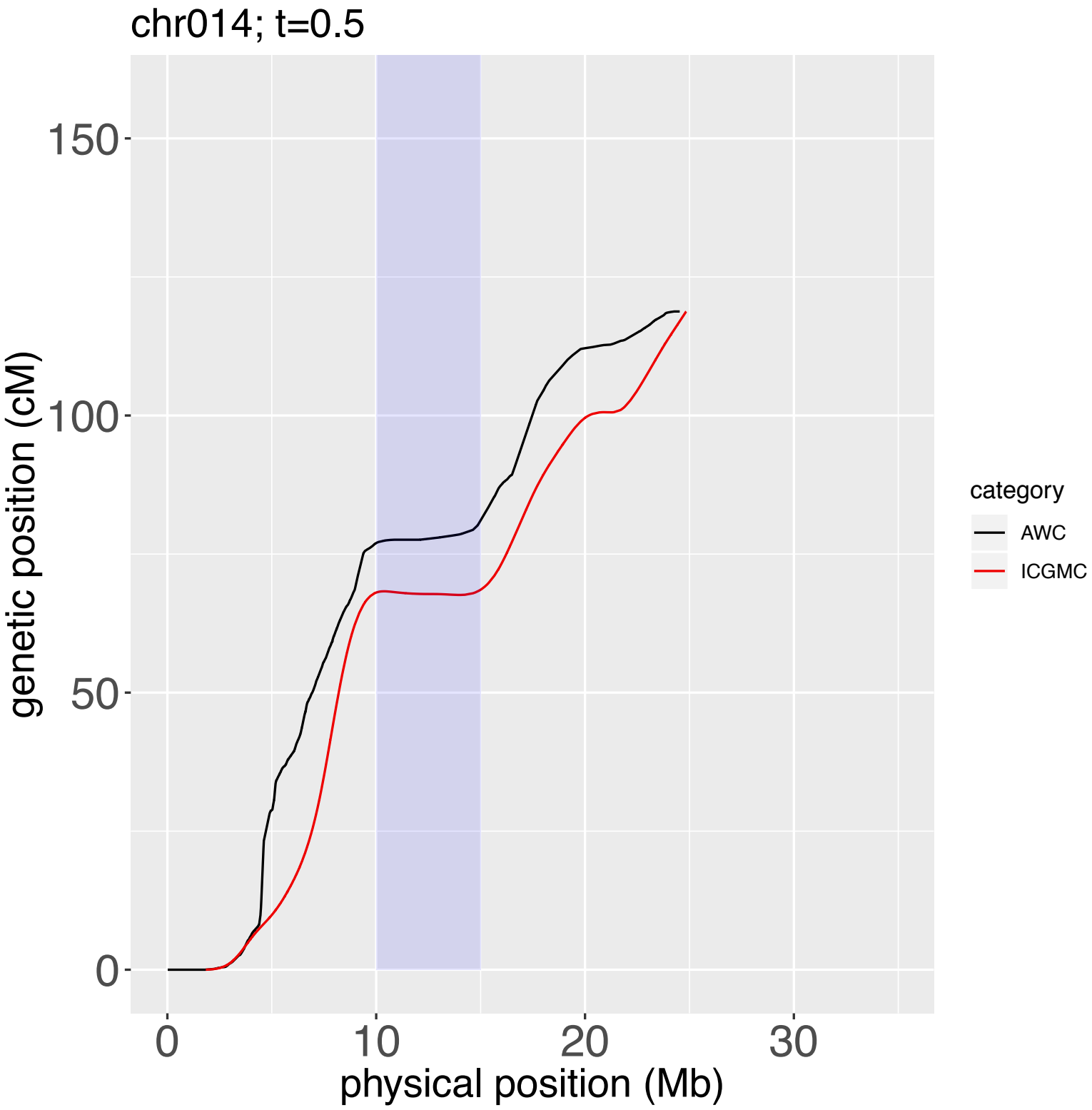

chr009; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

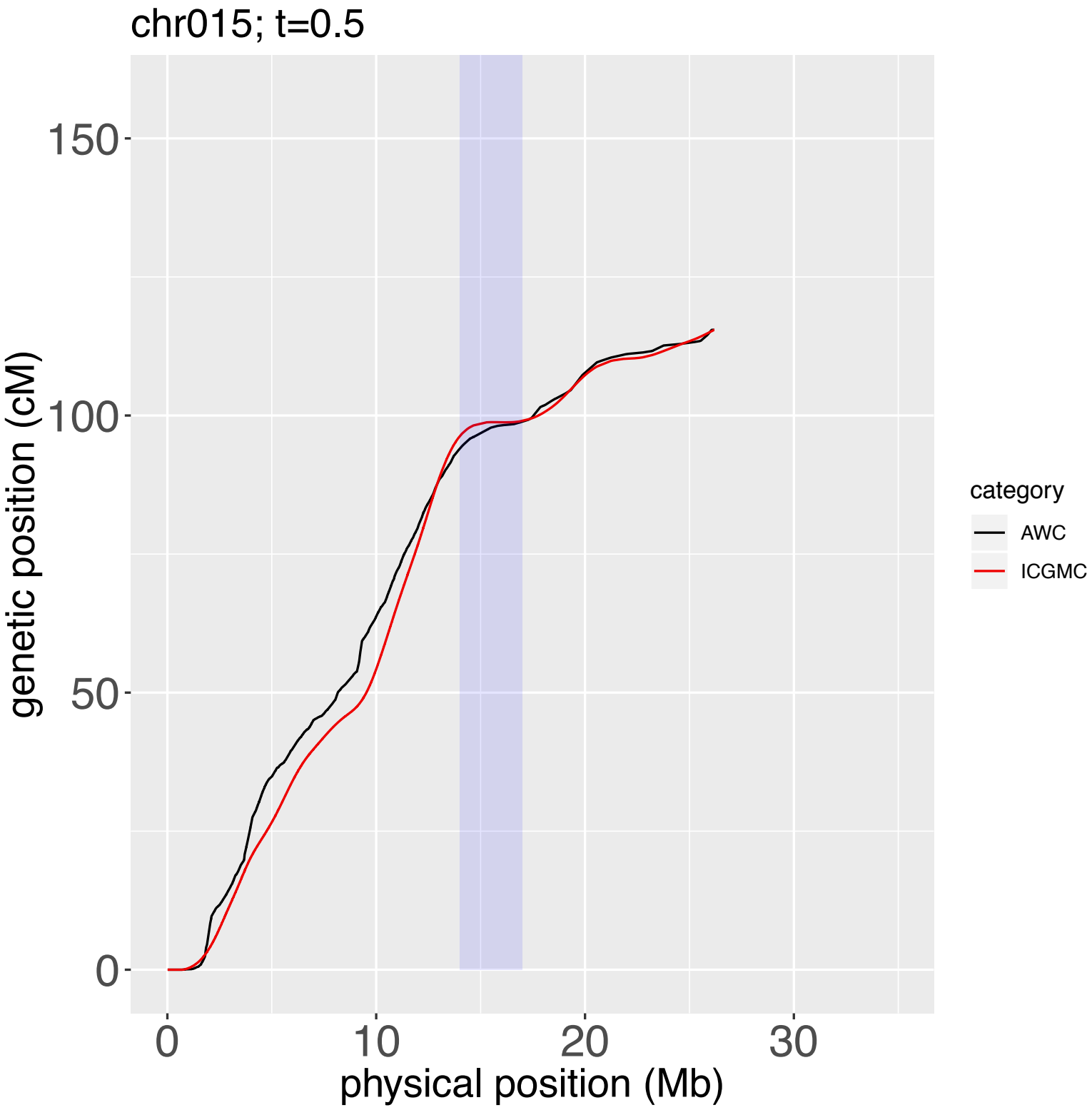

chr010; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

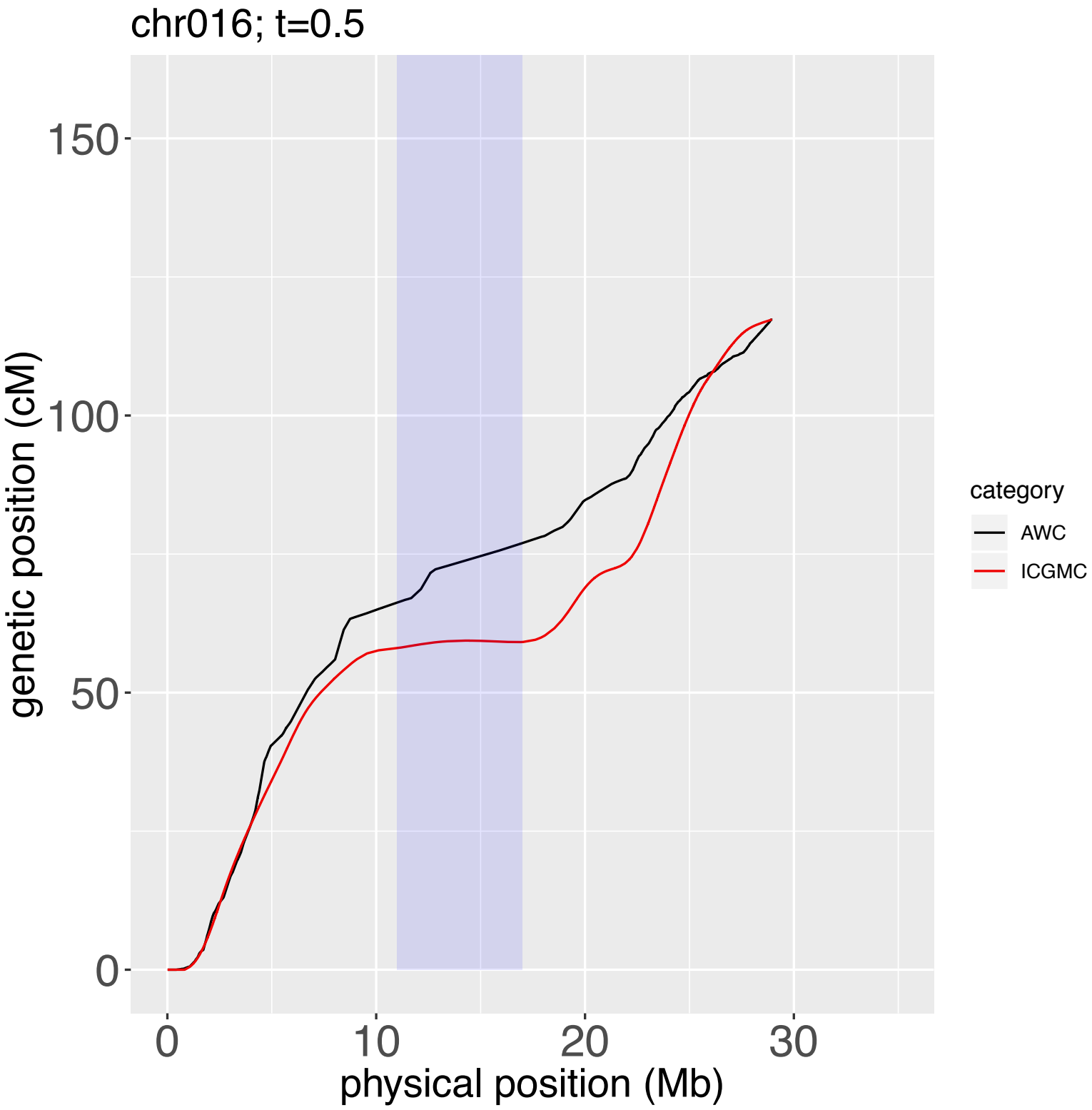

chr011; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

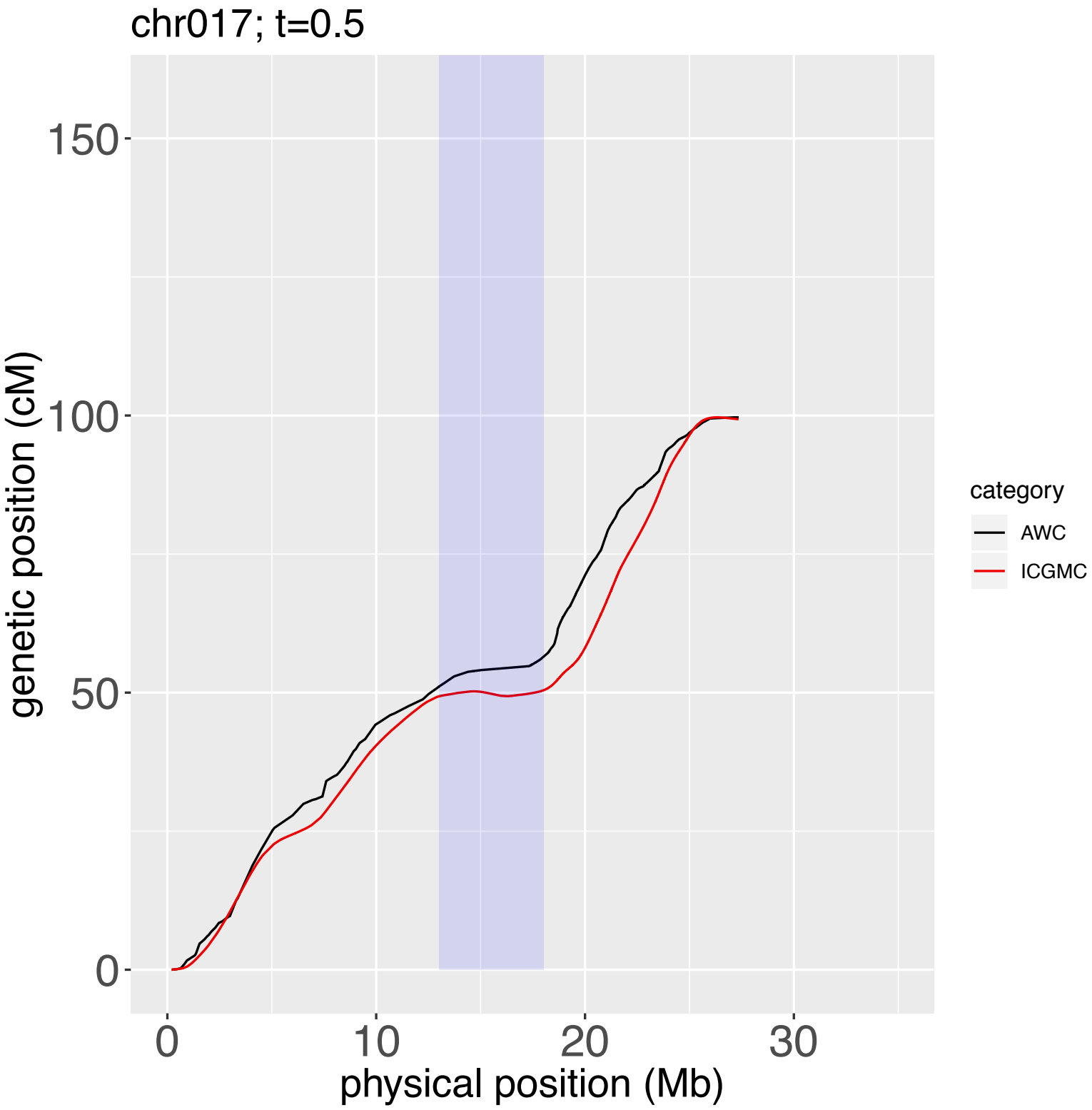

chr012; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

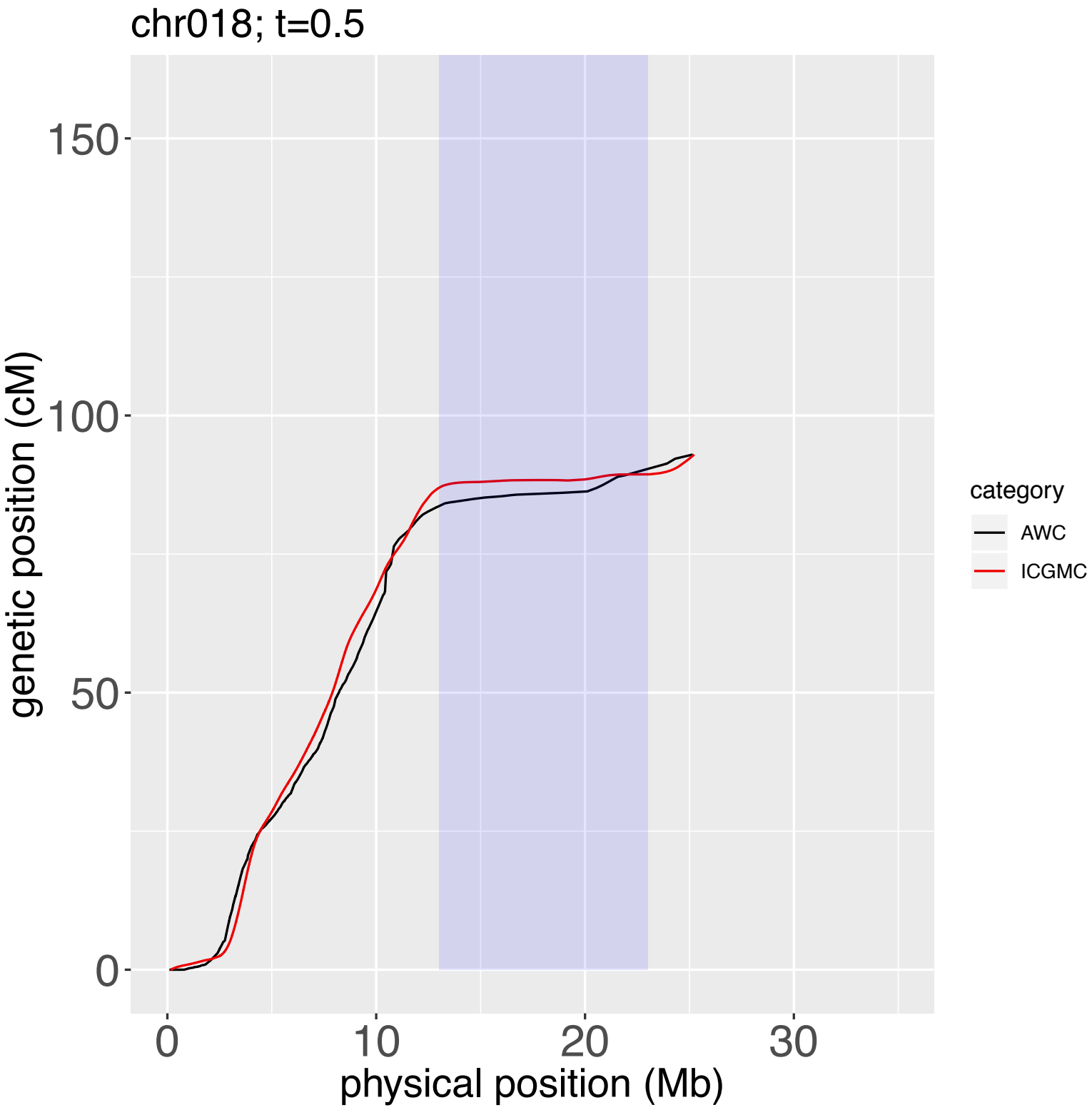

chr013; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC


chr014; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC


chr015; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

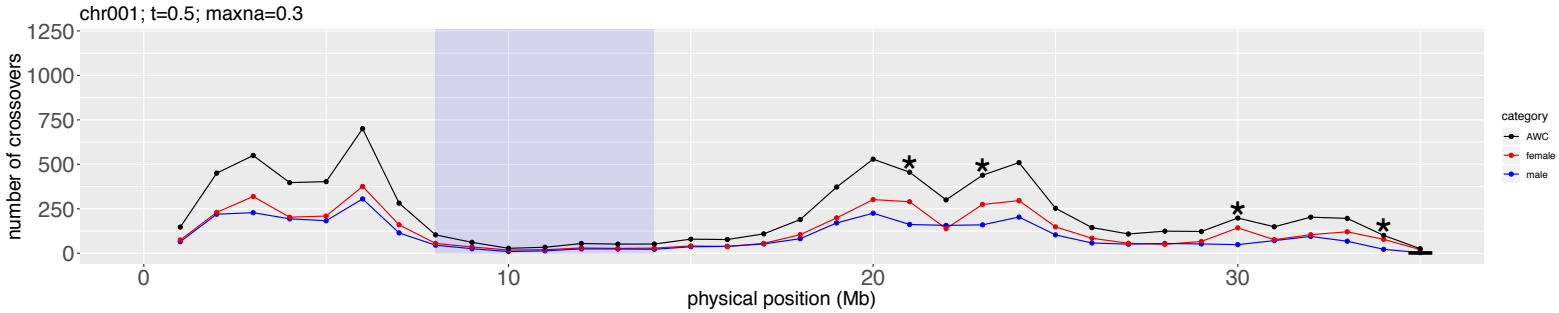

chr016; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

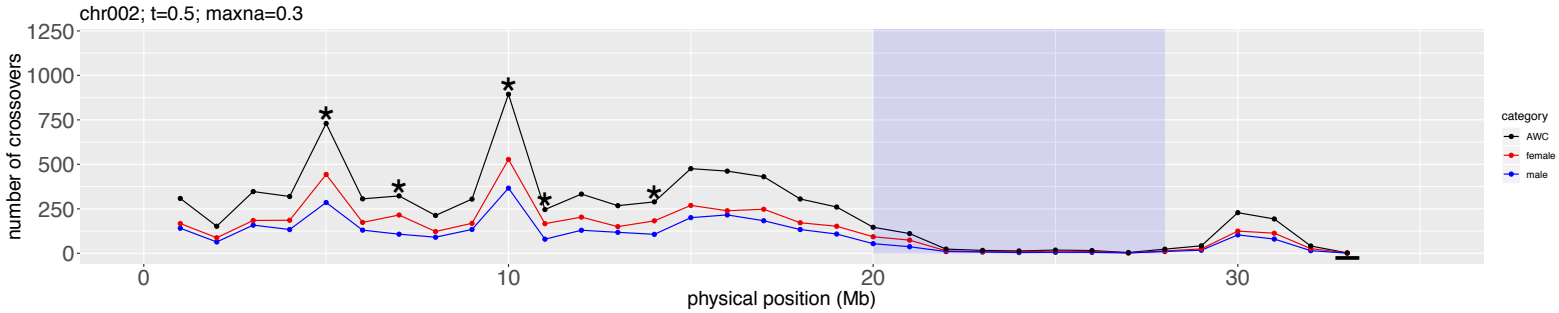

chr017; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

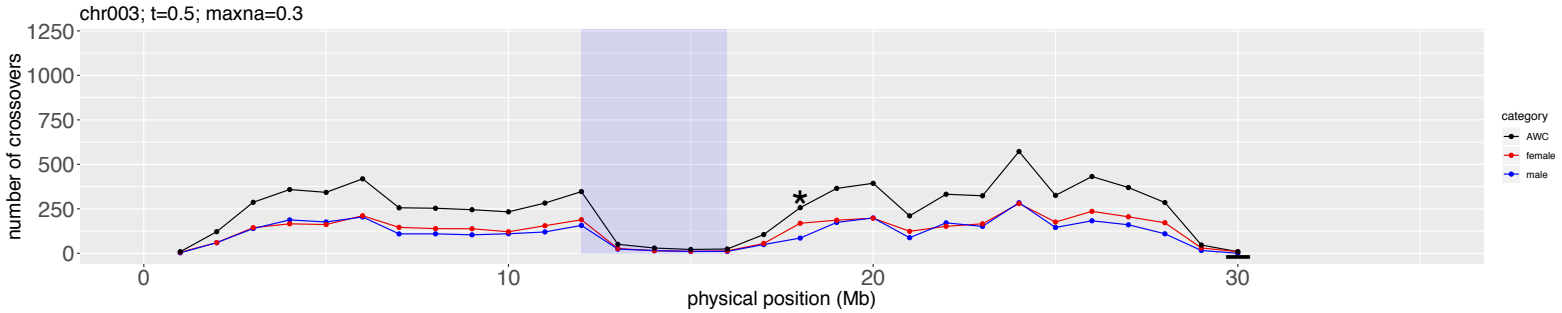

chr018; t=0.5
150
100
50
0
genetic position (cM)
0
10
20
30
physical position (Mb)
category
AWC
ICGMC

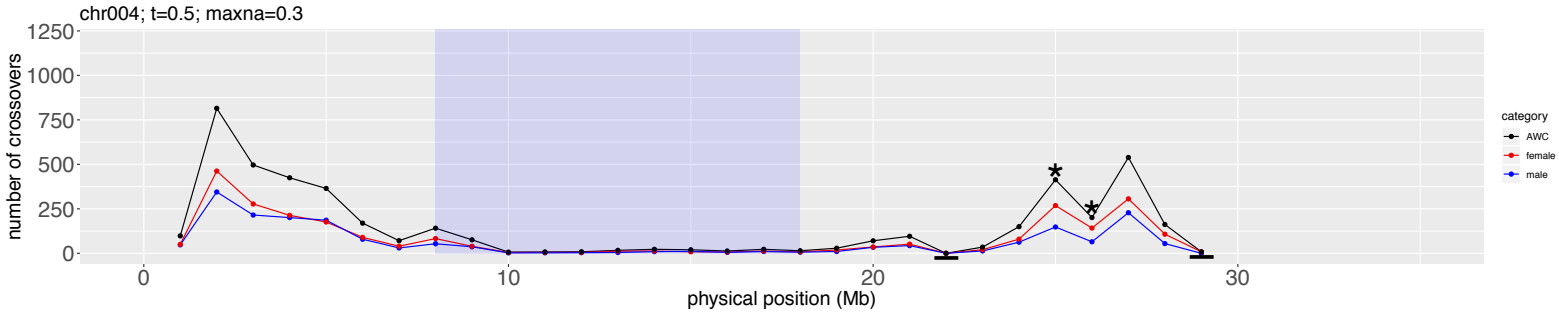
Appendix Figure 3.3

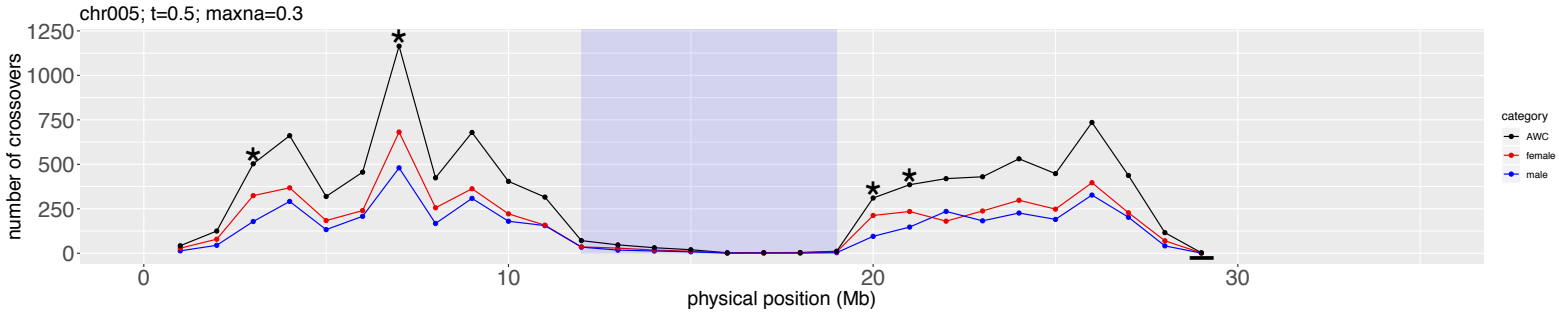
The 18 plots show the distribution of crossover events across cassava's 18 chromosomes for all meioses and female and males meioses, separately. We divided each chromosome into 1-Mb windows and plotted the number of crossovers falling within each interval for all (black), female (red), and male (blue) meioses. Asterisks show intervals with significantly different crossover counts between male and female meioses. Dashes represent cases where we could not perform the chi-square test because the expected frequency count for one or more classes was less than five. We did not test for statistical significance in the last window of any chromosome since the last window is shorter than 1-Mb (no chromosome is perfectly divisible by 1-Mb). These intervals are annotated with a dash. The centromere of chromosomes is shown in blue. We tested each interval at a significance level of α/n, where $\alpha = 0.05$ and $n = 506$.

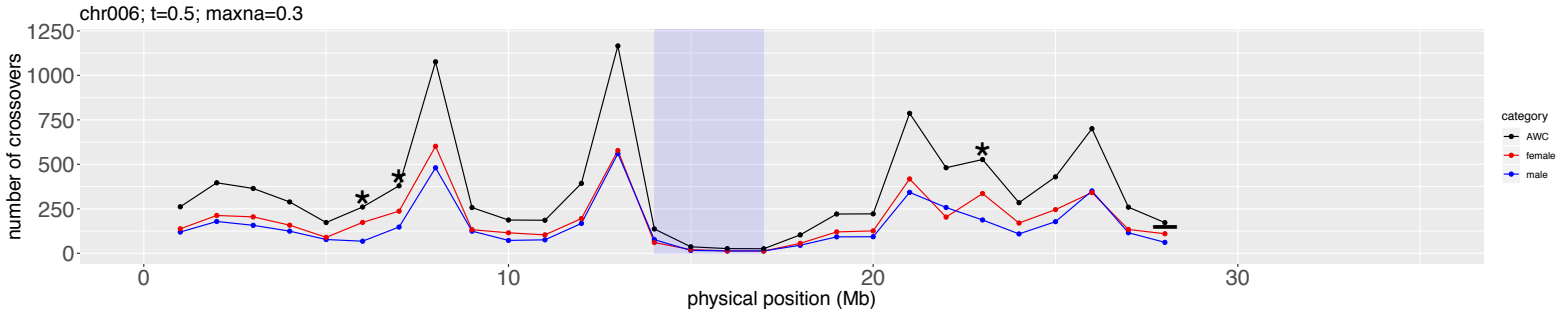

chr001; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male

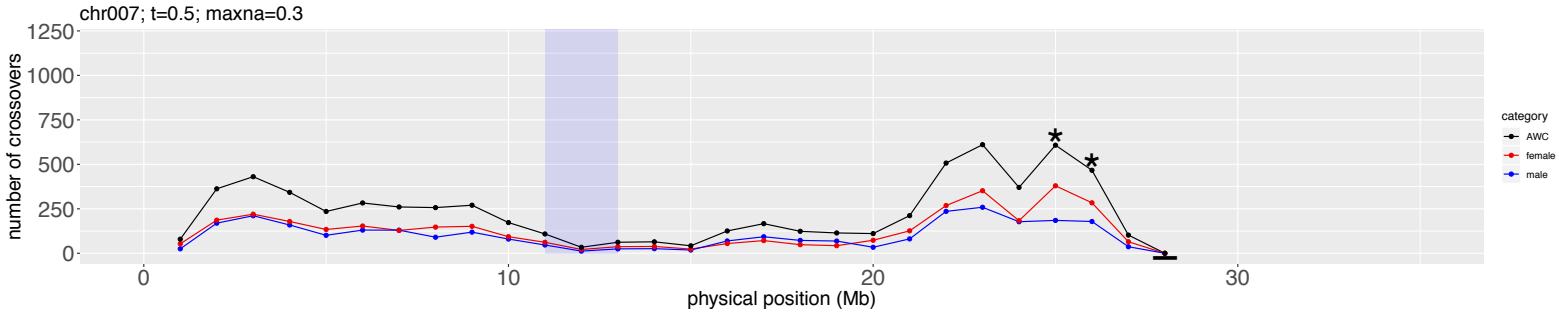

chr002; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male

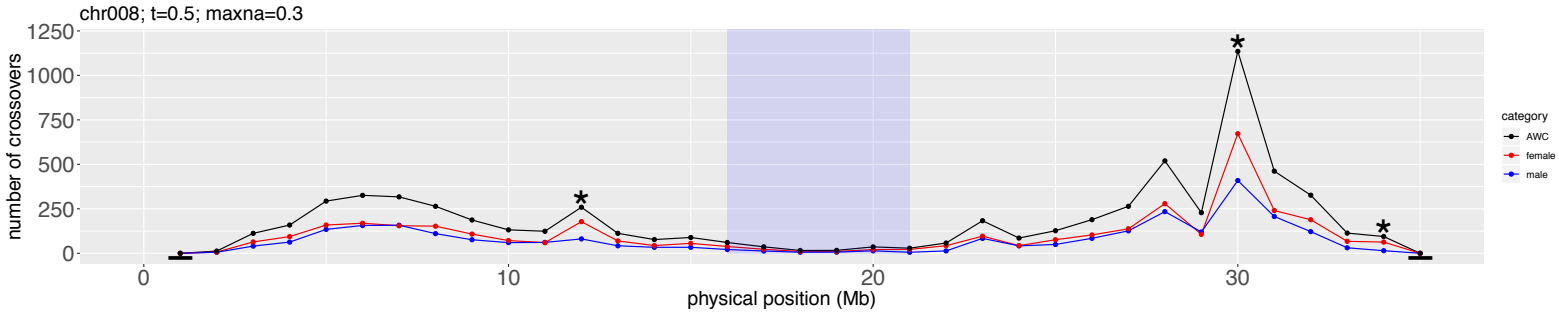

chr003; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male

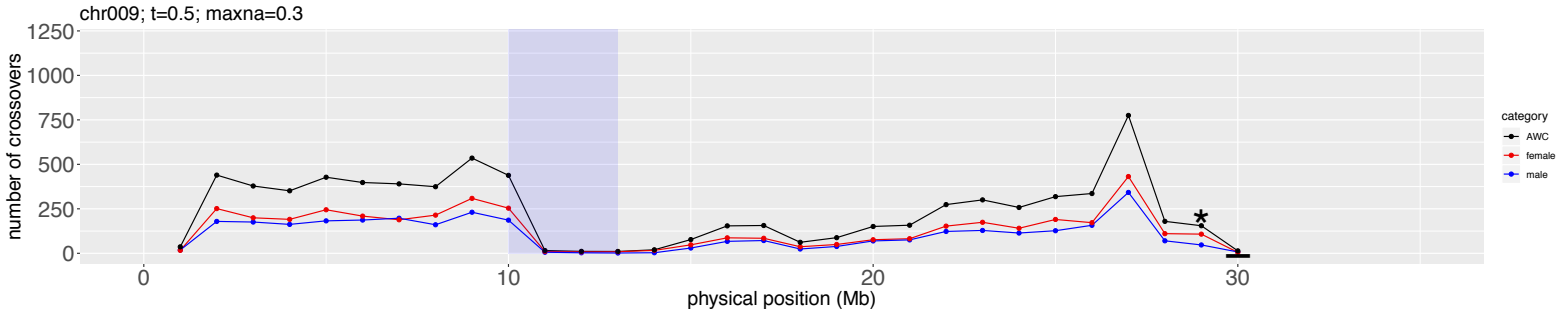

chr004; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male

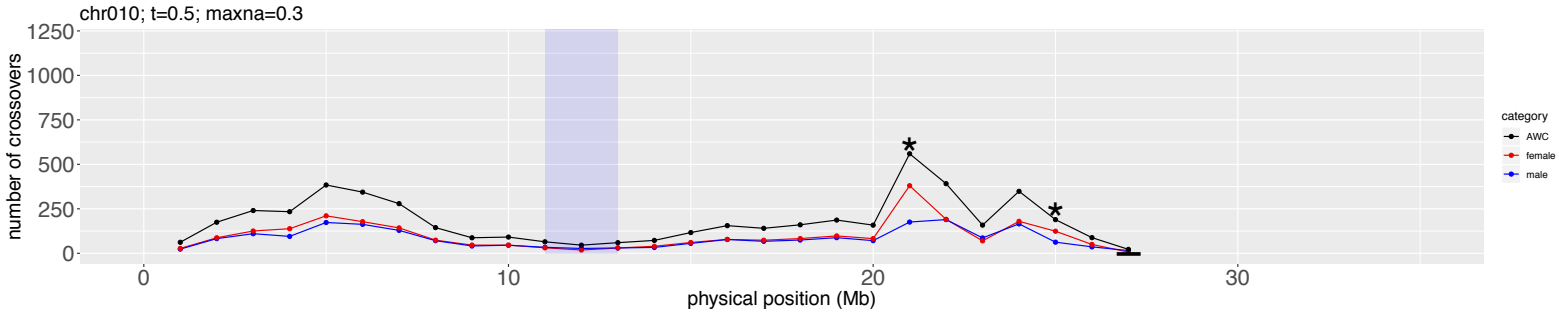

chr005; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male

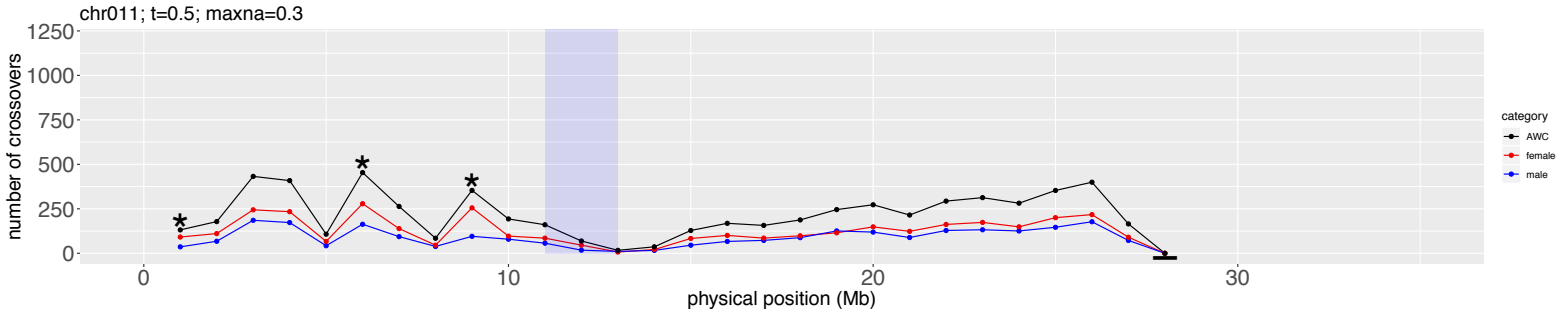

chr006; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male

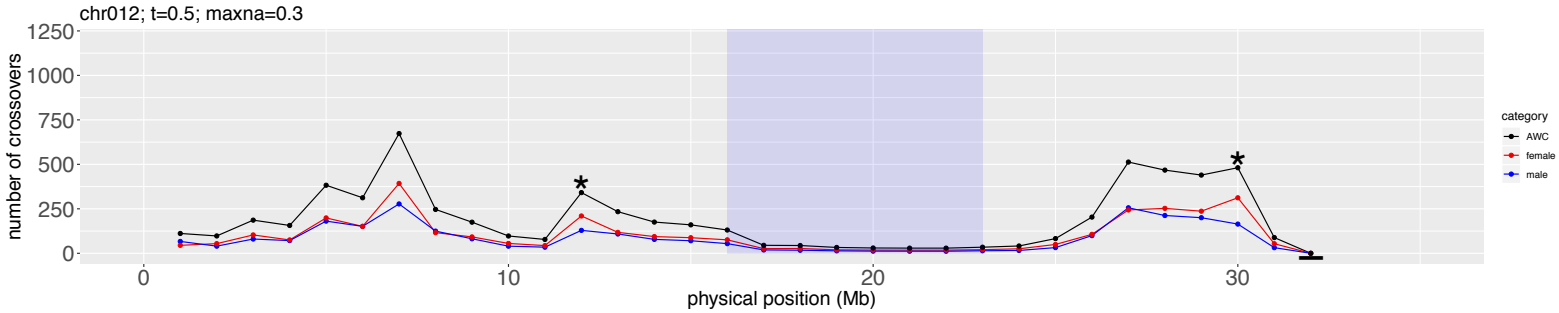

chr007; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male

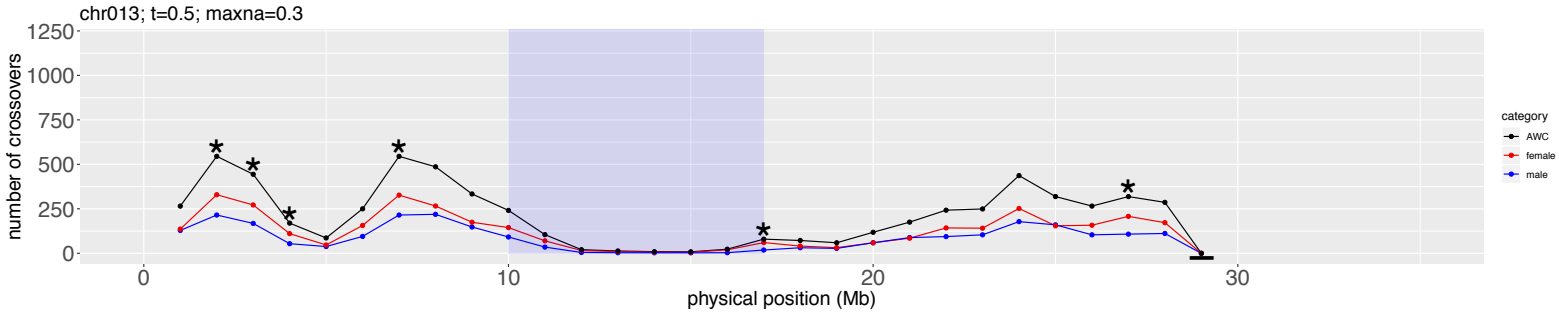

chr008; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male

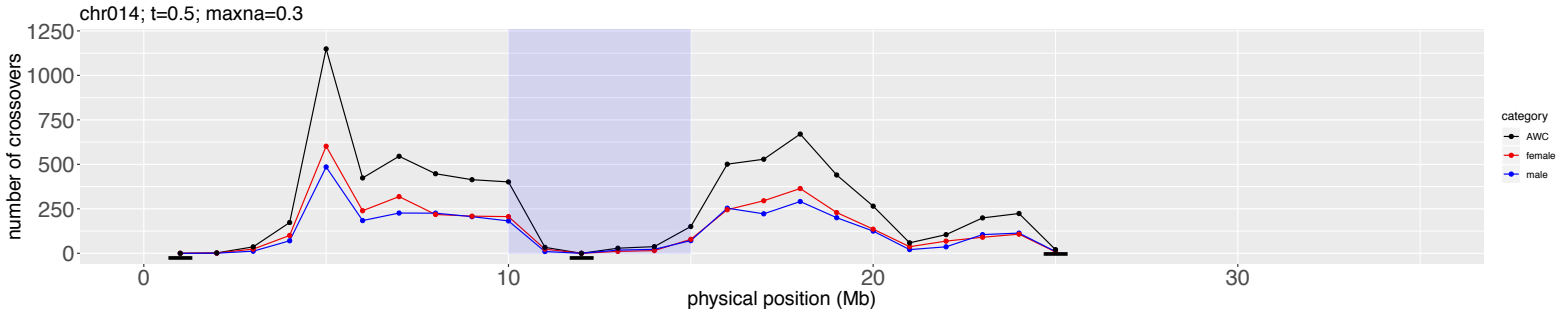

chr009; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male

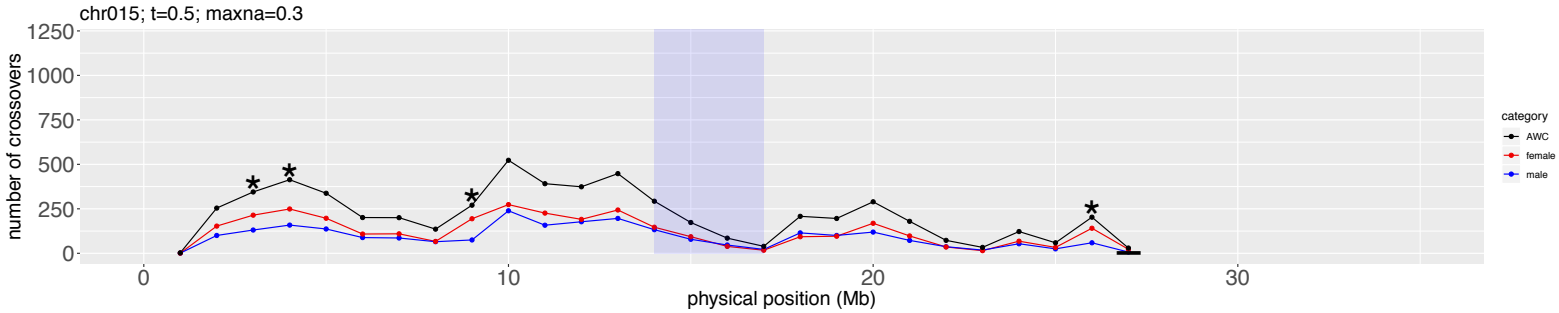

chr010; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male

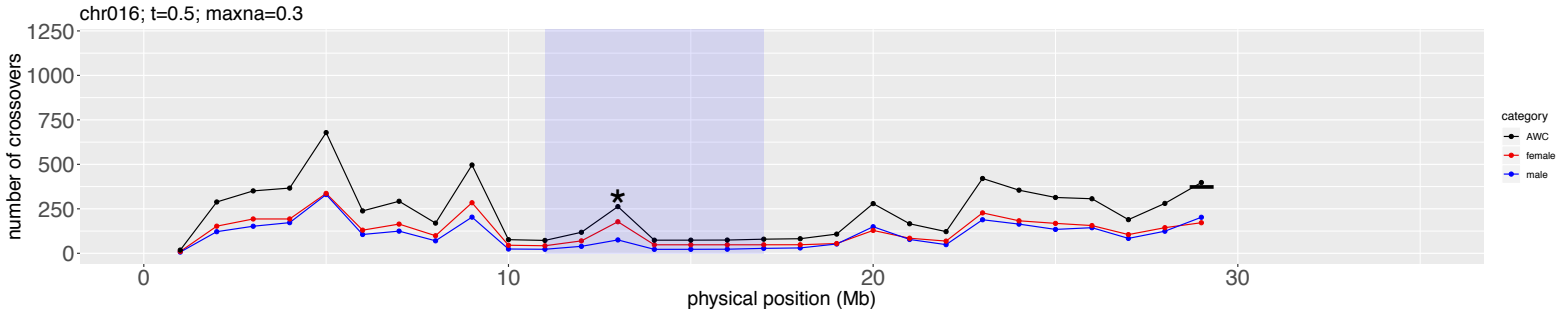

chr011; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male

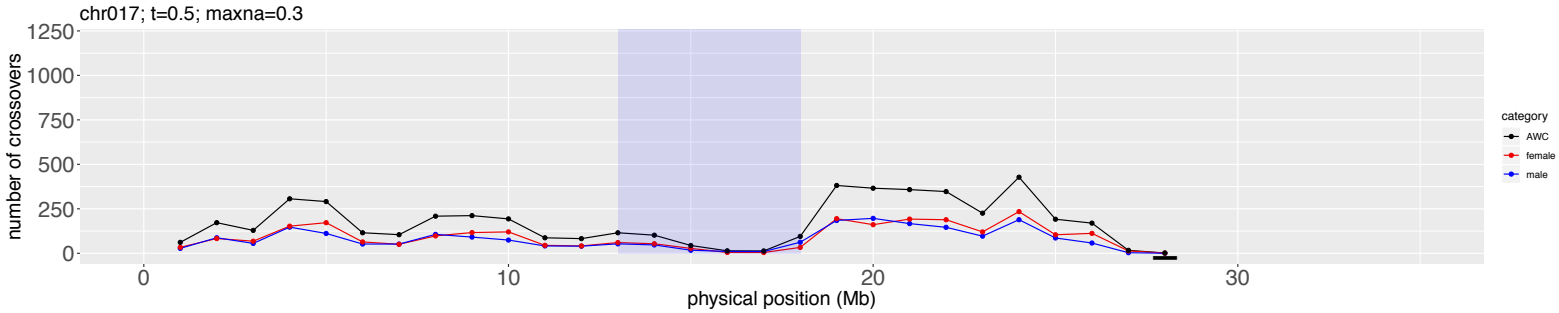

chr012; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male

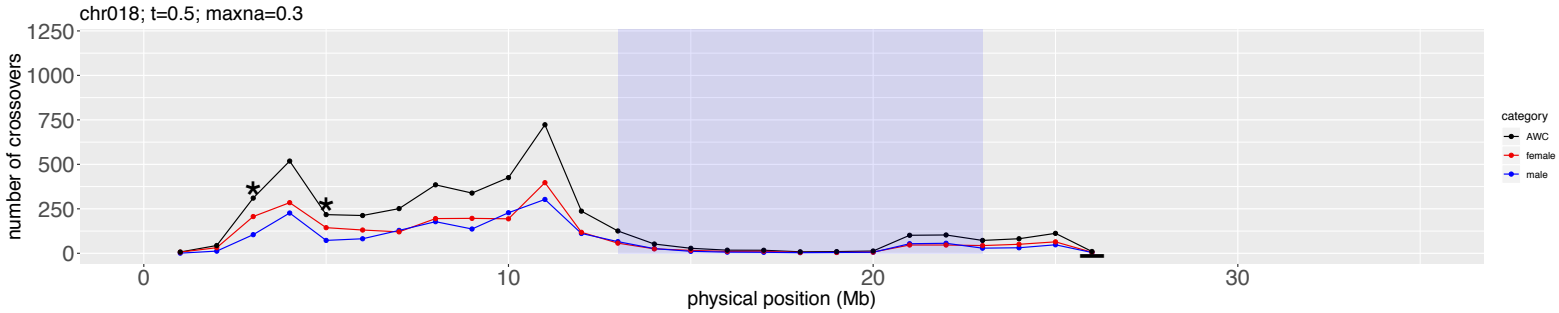

chr013; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male


chr014; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male


chr015; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male


chr016; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male


chr017; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male


chr018; t=0.5; maxna=0.3
number of crossovers
1250
1000
750
500
250
0
0
10
20
30
physical position (Mb)
category
AWC
female
male