UPPSALA
UNIVERSITET

Developing an Advanced Method for Kinship from Ancient DNA Data

Erkin Alacamli

Degree project in bioinformatics, 2023
Examensarbete i bioinformatik 45 hp till masterexamen, 2023
Biology Education Centre and Human Evolution Programme, Uppsala University
Supervisor: Torsten Günther

Abstract

The analysis of kinship from ancient DNA (aDNA) data has the potential to provide insight into social structures of prehistoric societies. Kinship analysis is gaining popularity as optimised wet-lab methods allow for studies with sample sizes on the level of whole cemeteries. However, the specifics of ancient DNA require different methods than what would be used for modern DNA. A common way is to use the sites that are identical-by-descent (IBD), however, detecting these is often a challenging task since it is not easy to determine whether a shared locus between two individuals is inherited from the ancestor or if another factor caused the similarity. Most methods used in the field are able to identify up to 2nd or 3rd degree relatives from aDNA data but do not distinguish between different types of relationship for the same degree, for instance not being able to differentiate between parent-offspring and full sibling-sibling relationship in first degree. The aDNA kinship methods often use either of window-based or single-site approaches, however, these two approaches have not been compared formally before in terms of effectivity and efficiency. In this work, READv2 is presented as a re-implementation of a popular kinship analysis method for aDNA studies with additional features such as accepting .bed files as input, which take up less space than the previous input type, plain-text .tped files. It is shown that the new version works more efficiently in terms of runtime. However, the memory requirements seem to be increased with the new implementation. Furthermore, a window-based approach is compared with the single-site approach of READv2, as well as varying window sizes, with benchmarked simulation data which contains approximately 700 individuals with known 1st degree, 2nd degree and 3rd degree relationships. According to the comparison, the sensitivity of the method does not vary between the approaches and different window sizes for high coverages. However, the single-site approach has been shown to be the superior one by a small margin for lower coverages. In addition to these, using the variance of non-shared alleles in windows along the genome has been used to implement a method to differentiate different first-degree relationships, parent-offspring and siblings. The method is tested with an independent dataset from the 1000 Genomes Project which shows that the proposed method is able to work with different datasets with varying sets of SNPs. Nevertheless, the first-degree classification method requires further analyses to determine the stress-point where the True Positive rates for both categories start to drop. Additionally, some necessary changes and decisions are required for READv2 to be a user-friendly method that can be used by other researchers. The preliminary release of READv2, including example data as well as instructions to install the necessary packages and to run the algorithm can be found in https://github.com/GuntherLab/READv2/releases/tag/READ.

READv2 – A Method to Reveal Ancient Kinship

Popular Science Summary

Erkin Alacamli

Studying ancient DNA (aDNA) can offer insights into the social structures of prehistoric populations. By examining the genetic connections within ancient populations, researchers can uncover valuable information about kinship and family relationships. The ancient kinship analysis field is rapidly developing, thanks to the advanced laboratory techniques that now allow for the study of entire ancient cemeteries. aDNA, however, presents unique challenges that requires different methods than those used for modern DNA analysis.

Identifying regions of the genome that are identical-by-descent (IBD), i.e. chunks of inherited DNA, is a common approach in kinship analysis. However, distinguishing the source of shared genetic markers between individuals, i.e. inherited from an ancestor or due to other factors such as mutation, can be challenging. Existing methods can detect relationships up to the second or third degree based on aDNA data. However, they cannot differentiate between various types of relationships within the same degree such as parent-offspring and sibling-sibling relationship in the first degree.

In this study, an improved version of a popular kinship analysis method called READ is introduced as READv2. READv2 offers enhanced features, such as accepting optimized input data and a new implementation of the previous algorithm. The new version shows better runtime efficiency, although there is a slight increase in memory demands. The study also compares two different approaches: the single-site approach of READv2 and a window-based approach, using varying window sizes. The tests on simulated data with known relationships showed that both approaches yield similar sensitivity and perform well with high coverage. However, the single-site approach outperforms the window-based approach, particularly with lower coverage. Additionally, a new method that uses the variance of mismatches along the windows within genome is developed to differentiate between parent-offspring and siblings. The tests using an independent dataset showed its effectiveness across diverse datasets with varying sets of genetic markers (SNPs). While this method shows promise, further analysis is needed to determine the stress-point, where the performance starts to decline. Furthermore, some changes and decisions need to be made to increase the user-friendliness of READv2, making it accessible to other researchers that are interested in using it.

Degree project in bioinformatics, 2023
Examensarbete i bioinformatik 45 hp till masterexamen, 2023
Biology Education Centre and Human Evolution Programme
Supervisor: Torsten Günther

Table of Contents

Abbreviations

aDNA	ancient DNA
DNA	deoxyribonucleic acid
IBD	Identity-by-descent
IBS	Identity-by-state
SNP	single nucleotide polymorphisms

1 Introduction

Kinship Studies in Ancient DNA

Analysis of kinship from ancient DNA (aDNA) data has the potential to provide insight into the social structures of prehistoric societies. It can give an idea of how closely related the individuals found in a Neanderthal cave (Skov *et al.* 2022), or in an Early Neolithic Tomb (Fowler *et al.* 2022) were, or it can be used to match different human remains such as a tooth or a bone to the same individual. Kinship analysis is gaining popularity as optimised wet-lab methods allow for studies with sample sizes on the level of whole cemeteries. However, the specifics of ancient DNA (low quality DNA, fragmentation, post-mortem damage) require different methods than what would be used for modern DNA. For instance, one of the widely used relationship inference methods, KING focuses on modelling the genetic distance between two individuals as a function of allele frequencies and kinship coefficient (Manichaikul *et al.* 2010). However, KING is not usually suitable for kinship estimation from ancient DNA due to ancient DNA samples being usually unable to yield diploid genotypes in which both alleles in a locus are available, because of low coverage and high missingness, and the absence of suitable reference sequences to infer allele frequencies. Due to the absence of diploid genotypes at both alleles, many methods ((Monroy Kuhn *et al.* 2018), (Fernandes DM *et al.* 2021), (Fowler *et al.* 2022)) assume that the genomes are pseudo-haploid which is inferring only the available or the most high-quality allele in a locus so that the individual ends up being represented by a haploid sequence (Barlow *et al.* 2020).

Most methods used in the field are able to identify up to 2nd or 3rd degree relatives from aDNA data but do not distinguish between different types of relationship for the same degree, for instance not being able to differentiate between parent-offspring and full sibling-sibling relationships in first degree. A common way to estimate the relatedness between the individuals is to infer the sites in the genome that are identical by descent (IBD), in other words, inherited by the same common ancestor. As the relationship between two individuals get closer, the proportion of sites that are in IBD rises.

The genome-wide proportions of sharing two chromosomes, one chromosome or zero chromosomes that are IBD can give information about the relatedness of the compared individuals (Popli *et al.* 2023). For instance, the proportions of these states for two siblings are 0.25, 0.5, 0.25; for parent-child relationship they are 0, 1, 0; and for unrelated individuals 0, 0, 1; respectively. However, detecting the sites in IBD is often a challenging task since it is not easy to determine if a shared locus between two individuals is inherited from the ancestor or if another factor caused the similarity by just comparing two sequences. Therefore, many

methods end up measuring the sites in IBS (identical by state), meaning the sites of interest that share the same allele without any knowledge on inheritance (Henden *et al.* 2018).

Kinship Estimation Methods

READ (Relationship Estimation from Ancient DNA) (Monroy Kuhn *et al.* 2018) is an example of the methods that infer family relationships for degraded samples, that can successfully infer up to 2nd degree relationships by using as little as 0.1x coverage per genome of individual pairs. READ requires only PLINK .tped and .tfam (Chang *et al.* 2015) files which includes SNP (Single Nucleotide Polymorphism) information for each individual for each locus (Table 1).

1	rs3094315	0.02013	752566	0	0	G	G	0	0	A	A
1	rs12124819	0.020242	776546	A	A	0	0	0	0	0	0
1	rs28765502	0.022137	832918	0	0	T	T	0	0	0	0
1	rs7419119	0.022518	842013	T	T	0	0	T	T	0	0

***Table 1*: An Example .tped file. 4 lines from a .tped file are shown here.** Each line represents a different site. The first 4 columns represent: Chromosome number, Variant identifier, Position in morgans (or centimorgans), Base Pair Coordinate. The rest of the columns show the allele information, each two columns belong to one individual, and missing information is represented as "0".

The workflow of READ (Figure 1) is quite simple when it is compared to its successors. The program checks for the proportion of P0 (non-shared alleles between individuals or in other words, the alleles that are not in IBS, i.e., 1 - (proportion of sites in IBS)), among the loci without any missing alleles present in any of the individuals. After that, a normalization value can be chosen from the genome-wide distribution of P0 values (median, maximum, or mean of the distribution as stated by the user) or be defined by the user. P0 values are then normalized with the normalization value. The normalization value represents the expected P0 value for a pair of unrelated individuals from the population, and the normalization takes place in order to eliminate the effect of the diversity of the population. After the normalization step, the relationship type is defined with the cut-off values shown in Figure 1.

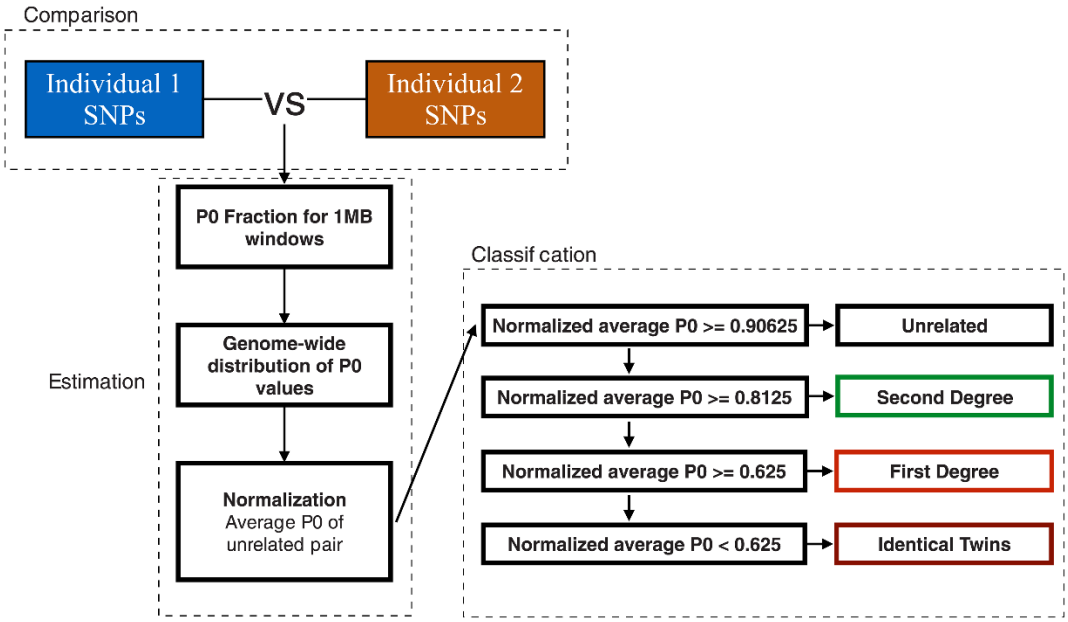


Figure 1: Workflow of READ. P0 stands for proportion of non-shared alleles in a window (Monroy et al. 2018)

Another newer method, TKGWV2 (Thomas Kent Genome-Wide Variants 2) (Fernandes DM *et al.* 2021), decreased the required coverage down to 0.026x. This method is also an upgrade to an older method which is reported to be effective in the range of 0.04X – 0.1X coverage (Fernandes D *et al.* 2017). Compared to READ, TKGWV2 requires same files and information, such as files with aligned reads per individual (BAM files), list of SNPs for genotype calling, with the addition of population allele frequencies of the same SNPs. However, if the input files are already in PLINK .ped format, then, only allele frequencies are needed. The main workflow of TKGWV2 includes genotype calling and creating individual pseudo-haploid PLINK .ped text files. As mentioned above, if PLINK files are included in the analysis, this first step can be skipped. Afterwards, the method identifies the overlapping variants for each pair of individuals, the corresponding allele frequencies are extracted and a transposed PLINK text file(.tped) is created, which is then used to calculate pairwise relatedness by taking the average of Queller and Goodnight's relatedness estimator (Queller & Goodnight 1989) expressed by the following formula (Oliehoek *et al.* 2006) for each pair at locus *l*, where *xy (*and *yx)* is the pair of individuals x and y, *I* is the identity for alleles *ab* for individual x, and *cd* for individual y. Finally, p_a and p_b denotes allele frequencies for these alleles:

$$r_{xy,l} = \frac{0.5*\left(I_{a_c}+I_{a_d}+I_{b_c}+I_{b_d}\right)-p_a-p_b}{1+I_{a_b}-p_a-p_b}$$

More recently proposed method KIN (Popli *et al.* 2023) is reported to achieve up to 3rd degree relationship classification, and differentiation between parent-offspring and sibling pairs for first degree relationships, while requiring as low as 0.05x sequence coverage by the utilization

of a Hidden Markov Model (HMM) which is a finite model that is composed of states and state-transition probabilities that connect these states (Eddy 1996). KIN fits a model called *KIN-HMM*, the goal of which is to infer the relationship between a pair of individuals by incorporating the patterns of common IBD states along the genomes. For this purpose, KIN divides the genomes of the pair into windows, and fits the data for each pair of individuals to the KIN-HMM model for each considered relatedness case (e.g., parent-offspring, siblings, grandparent-grandchild and so on), and the case with the highest likelihood is then chosen to classify the relationship between the pair.

KIN requires information on the number of overlapping sites for windows where both individuals have information available, the number of pairwise differences at these sites, and the probability of runs of homozygosity (ROH) in the windows where individuals have the same long allele sequence. By default, this value is obtained from another HMM called *ROH-HMM*. Additionally, KIN provides another python package, *KINgaroo* (Popli *et al.* 2023), that generates the input files for the abovementioned models from *.bam* files, and additional models that correct the data for contamination and inbreeding in order to improve classification accuracy.

As explained in the abovementioned paragraphs, while TKGWV2 uses a single-site-approach, READ and KIN use window-based approaches that compare windows of different sizes in the genomes of two individuals, similar to the study of Fowler et al. 2022. The difference with this method (Fowler *et al.* 2022) is that they compute the mismatch rates by dividing the genome in windows of size 5Mb, and the relatedness coefficient is computed as a function of mismatch rate of the pair and expected mismatch rate of an unrelated pair from the same population. The latter value is estimated from the genomic data published by the studies that targeted the same population (Brace *et al.* 2019), (Olalde *et al.* 2018). Afterwards, for the differentiation of different types of first-degree relationships, they have checked the DNA sharing patterns by computing allelic mismatch rates across sliding windows of size 20Mb with step sizes of 1Mb. The presence (sibling pairs) and absence (parent-offspring) of regions with zero or two chromosomes shared was used to determine the type of first-degree relationship. Moreover, they replicated the analysis with another method called ngsRelate (Hanghøj *et al.* 2019) in order to test the validity of the results that they received from their own method. ngsRelate goes over every possible genotypic configuration and assigns a probability based on their genotype likelihood, and hence, considers the possible uncertainty of the genotypes.

All these mentioned studies have been used or are still being used in different kinds of research areas that involves kinship analysis from aDNA. Although they report high performances in terms of the power to estimate relationships, they all have their own advantages, disadvantages, and preferred scenarios. For instance, methods like TKGWV2 need extra information about the samples, such as population allele frequencies or genotype likelihoods and so on. Therefore, if the population is unknown, or there is not much information and/or studies done about the population of interest, then the results of the

analysis might be noisy and unreliable. As mentioned before, READ is a quite basic and straightforward method, but it does not correct for contamination or inbreeding which might affect the performance (Popli *et al.* 2023). On the other hand, KIN corrects for these factors and additionally, the hidden part is modelling the unobserved diploid genotype unlike the other methods, however, the preparation of the data is more complex and time-consuming. While ngsRelate and lcMLkin (Lipatov *et al.* 2015) which utilize genotype likelihoods to infer relationships, are used widely used in aDNA research. They are not specifically built for low coverages that is usually the case for ancient samples. Besides the methods mentioned above, there are some more recent developments that seem to be promising such as ancIBD (Ringbauer *et al.* 2023) and correctKin (Nyerki *et al.* 2023), yet they need a lot of data to operate.

Aim of The Project

The goals of this project include reimplementation of READ (Monroy Kuhn *et al.* 2018) in Python3 in order to avoid compatibility and support issues that may arise from the fact that Python2 is deprecated and adding new features such as being able to accept .bed PLINK files as input in order to save space in the hard drives. Although the mentioned studies use either a window-based or single-site approach, these or different window sizes have not been formally compared in terms of performance and effectivity before. Therefore, the effect of different window sizes in window-based approach, and single-site approach on the performance of the method is compared. As illustrated in Marcus et al. 2020, in first degree relationships, the difference along windows in parent-offspring relationships varies less than sibling-sibling relationships, since an offspring always acquires one copy of chromosomes from their parent while siblings can share zero, one or two chromosomes at a given locus. The possibility of using this knowledge to differentiate between different kinds of first-degree relationships is investigated. Afterwards, this method is tested with an independent dataset from the 1000 Genomes Project.

2 Materials and Methods

2.1 READ reimplementation in Python3

READ (Monroy Kuhn *et al.* 2018) is written in Python2 and R, in a way that the R script is called in the Python script to carry out some analyses including normalization of P0 values, calculating average normalized P0 value per pair, standard error calculation of windows and so on. A detailed description of the READ workflow can be found in the Introduction section. The first step of the project was to translate the READ script to Python3, in order to update the script, and increase efficiency, portability, and to avoid possible version conflicts. The

parts of READ previously written in R were implemented using the Pandas (McKinney 2010) library in Python3. Furthermore, with the reimplementation, the feature of being able to read PLINK .bed, .fam and .bim files was added using the Python *PLINKIO* (Frånberg 2021) library. In order to avoid loops and improve the runtime of the method, the pairwise comparison was implemented with the NumPy (Harris *et al.* 2020) library. The performance of the new implementation was tested and compared with the old implementation in terms of runtime and memory usage with the Linux *time* command. The test data (Mathieson *et al.* 2015) includes 14 individuals from the Srubnaya culture genotyped at 1.2 million SNPs. The effect of halving the number of individuals (with the Plink command –thin-indiv 0.5) and the number of SNPs (with the Plink command –thin 0.5) was investigated in the test as well.

2.2 Simulated data with known relatedness

The next step after the reimplementation was to create a benchmark using simulated data with known relationships to test the performance of READv2 in terms of sensitivity and false positive rates. The simulated NGS data was created by Sevval Akturk, Merve Nur Guler, Igor Mapelli and Kıvılcım Vural from the Comparative and Evolutionary Biology Lab in Middle East Technical University (METU), with the PED-SIM software (Caballero *et al.* 2019). The founders of the pedigrees were created from scratch by estimating allele frequencies from modern-day Tuscan individuals from the 1000 Genomes Project (Auton *et al.* 2015). The alleles were randomly drawn from them, in order to eliminate the background relatedness between relatively close pairs, and for each pedigree a pair of founders was chosen randomly. 96 pairs of first-degree relationships, and 144 pairs of second-degree relationships with distinct founders (i.e. none of the individuals in first-degree relationships were related to the individuals in second-degree relationships) were used in testing. The genome-wide autosomal genetic data of all pairs were down sampled to the same 200,000 randomly selected SNPs from the Tuscan population from 1000 Genomes Project Phase 3.

Next, next-generation sequencing data for the pedigrees were simulated as ancient samples by cutting the reads in variable lengths to mimic the distribution of actual ancient reads and adding post-mortem DNA damage with Gargammel software (Renaud *et al.* 2017). These ancient reads were later mapped to the hs37d5 human reference genome using BWA (Li & Durbin 2009) with “aln” option. The reads mismatched to the human reference genome were eliminated with a cut-off of 10%, moreover, the remaining reads were trimmed from both ends to remove C-to-T (or G-to-A) substitutions that stem from post-mortem DNA damage.

2.3 Sub-sampling and benchmarking

Due to the large size of the simulated dataset (n = 696), in order to lower the memory and runtime requirements of the analyses, the dataset was divided into groups of 70. Individuals from the same relatedness group, such as parent-offspring trios, were kept together in the groups using the PLINK –keep-fam command. The READ normalization value was

calculated as the median of all the samples, however since this value was not significantly different than group medians (Table 2), the default normalization method and value was inferred for each group separately and used in successive analyses.

Coverage	Global	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10
5X	0.2447	0.24635	0.2457	0.2445	0.2435	0.2454	0.24485	0.2425	0.2425	0.2455	0.2429
1X	0.2439	0.24635	0.2435	0.2423	0.24385	0.24245	0.24504	0.244	0.2428	0.2431	0.243
0.5X	0.2444	0.2468	0.2439	0.241	0.2469	0.2446	0.2445	0.2445	0.24255	0.241	0.2415
0.1X	0.2509	0.2546	0.2368	0.2545	0.241	0.24195	0.25075	0.2562	0.231	0.248	0.2439
0.05X	0.2549	0.256	0.257	0.251	0.238	0.2439	0.24395	0.2461	0.2431	0.2143	0.2578
0.01X	0.2654	0.305	0.15	0.213	0.22	0.266	0.13245	0.219	0.2538	0.1788	0.4

Table 2: Global and per-group median values of the mean P0 values for each coverage. The values were recorded after the analysis was done for the window size of 1Mb (default) for each group. The global value was recorded by taking the median of concatenated mean P0 values of each pair from each group.

In order to test the performance of READ for different coverages, the original simulation data (5X coverage) was down sampled to 1X, 0.5X, 0.1X, 0.05X, and 0.01X coverages with SAMTOOLS view -s (Danecek *et al.* 2021). Then, true positive (TP), false positive (FP) and false negative (FN) values of the classifications for each coverage in default window size (1Mb) was acquired. Afterwards, in order to see how window size affected the results and to compare the window-based and single-site approaches, the power of the method, i.e. sensitivity (TP/TP+FN), and the proportion of false positive unrelated matches to all unrelated matches, i.e. the pairs in first or second degree relationships but classified as unrelated, were acquired for each coverage and window sizes (100Kb, 1Mb, 5Mb, 10Mb and 20Mb).

For further testing on real data, empirical data from the 1000 Genomes Project (Auton *et al.* 2015) was used. The autosomal Illumina Omni2.5M chip high density genotype call SNP array data consists of 2368 individuals from 15 different populations with 2,458,861 SNPs. In order to further filter the data, first, the number of parent-offspring and sibling pairs for each population was calculated (Table 3). Then, the populations with the most sibling pairs and a sufficient number of parent-offspring pairs, namely ASW (African Ancestry in Southwest US) and CHS (Southern Han Chinese, China), were selected for the further steps. The chosen populations were separated into different .bed files with PLINK –keep-fam option, and later down sampled to 50%, 25%, 10%, 5% and 1% of the SNPs with PLINK –thin option. The data was from modern samples; therefore, the samples were diploid and included heterozygous sites. However, this could be problematic for READv2 since it assumes that the data is pseudo-haploid. In order to solve this issue, the samples were made pseudo-haploid by randomly selecting an allele at each position with a python script (Appendix A).

Pop. Name	PO Pair	Sib. Pair
ASW	46	8
CDX	2	NA
CEU	1	NA
CHS	105	8
CLM	69	NA
GBR	2	1
GIH	NA	NA
IBS	100	NA
KHV	41	1
LWK	5	NA
MXL	60	3
PEL	70	1
PUR	66	NA
TSI	1	NA
YRI	112	4

Table 3: The Number of Parent-Offspring and Sibling Pairs that are Found in Each Population. PO Pair and sib. Pair stands for parent-offspring and sibling pairs respectively.

3 Results

3.1 Re-implementation of READ

READ was reimplemented in Python3, and PLINKIO was used to add the ability to read PLINK .bed files. Figure 2 shows the runtime and memory usage of the new version run with the full example data compared to the old implementation, the new version with SNP thinned data (to 50% of the SNPs) and the new version with individual thinned data (to 50% of the individuals). After the new implementation of READ, the runtime and the memory usage seem to be worsened due to the fact that the entire genotype data is stored in the memory while READv1 processed the data line by line and printed intermediate results to a file. It can also be deducted that the number of SNPs seems to be in a linear relationship with both the runtime and the memory usage, while the number of samples (i.e. individuals) seem to be in a more quadratic relationship due to the pairwise comparisons.

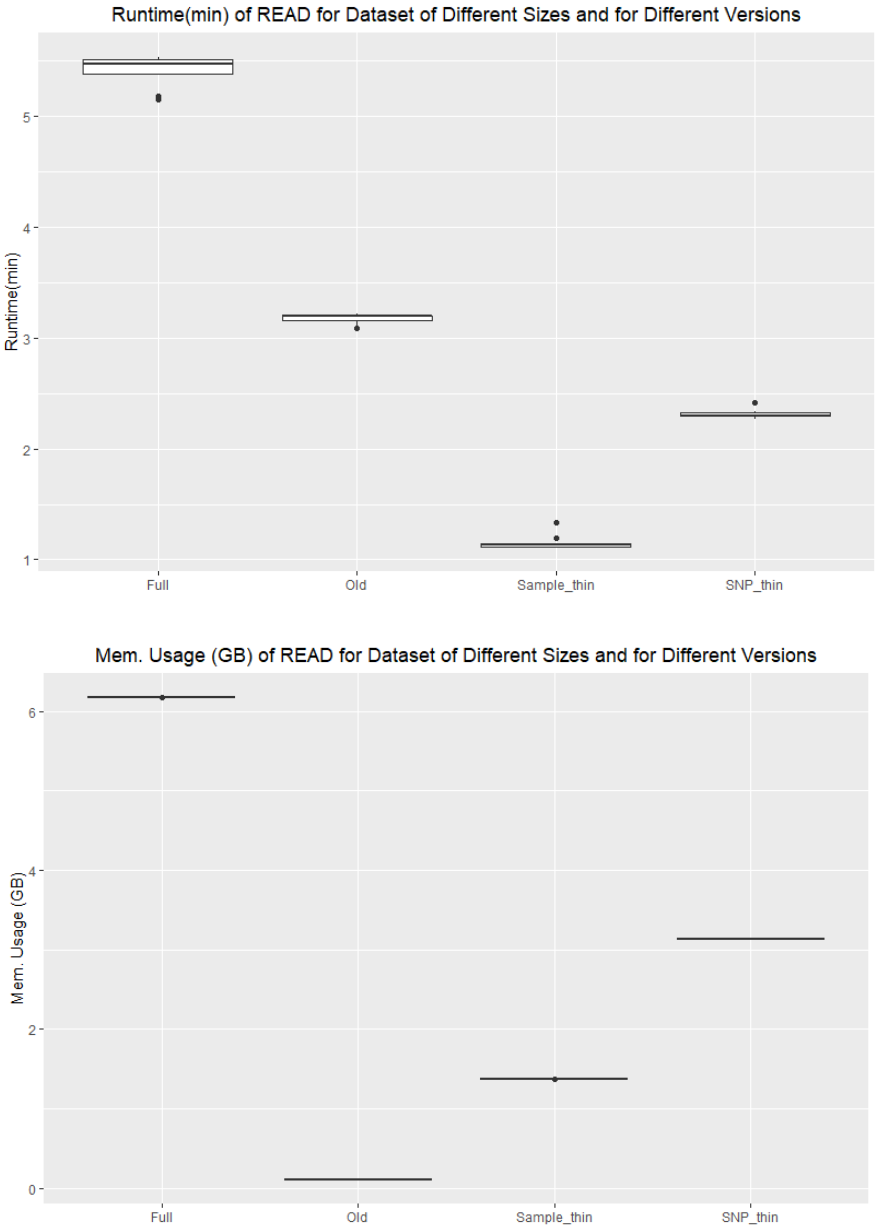


Figure 2: Runtime (in minutes) and memory usage (in GB) of READ with the full example data in different versions (labeled as full, i.e. the new implementation, and Old, i.e. the previous implementation), and the new version with half of the individuals (labeled as Sample_thin) and half of the SNPs (labeled as SNP_thin). Full data consists of 14 individuals, and 1.2 million SNPs. READ was run 10 times in each case in order to eliminate the eventual effect of processor and some other random effects.

In order to decrease the runtime, the pairwise comparisons were implemented as Numpy comparisons (this version is called as READv2 in the successive parts of the text) instead of pairwise comparison in loops. Figure 3 shows the runtime and memory usage of the new implementation with above-mentioned data, as well as a single site, not window-based implementation. After the modification, READv2 seems to work more efficiently in terms of runtime, compared to READv1. However, although there is quite an improvement, it still is lacking in terms of memory usage when compared to READv1. The same relationships of

sample thinned data and SNP thinned data with full example data seem to hold in this case, as well.

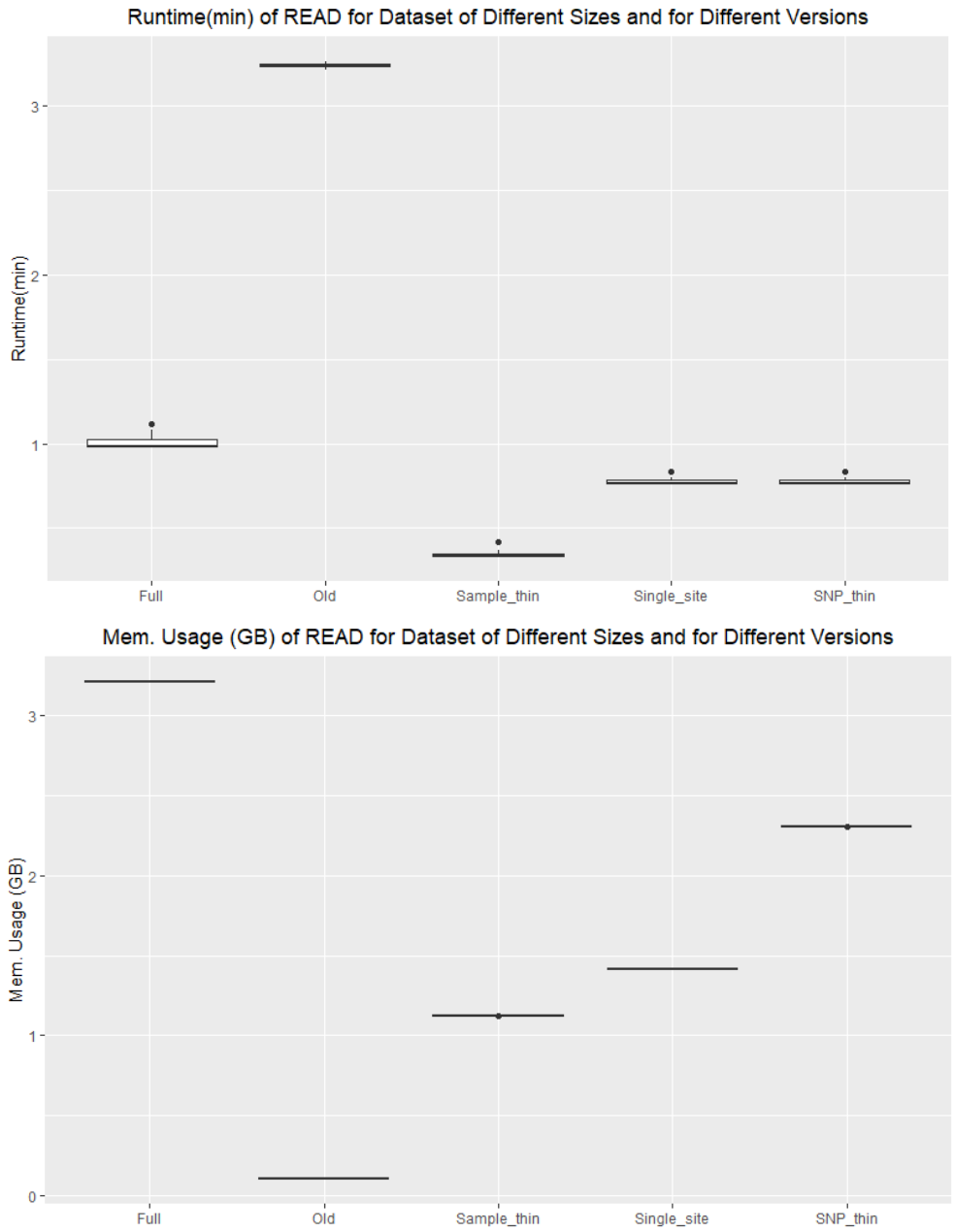


Figure 3: Runtime (in minutes) and memory usage (in GB) of READ for the newly implemented version. Modified READv2 ran on full example data compared to the old(i.e. READv1) version, single site version with full data. As explained above, the effect of halving SNPs (SNP_thin) and halving the number of individuals (Sample_thin) is shown as well. READ was run 10 times in each case in order to eliminate the eventual effect of processor and some other random effects.

3.2 Window size and Single-Site Approach Comparison

The simulated data with known relationships was then used to analyse the comparison of the effects of a single site approach and the previously used window-based approach, as well as the effect of varying window sizes. Figure 4 shows the power of READv2 under different

window sizes and different coverages. The sensitivity seems to be the same for high coverages (i.e. 5X, 1X for both degrees and 0.5X for first degree relationships), first degree classification can be observed to perform better than second degree classification in each coverage and window size. As the coverage goes lower, the single site method seems to be the best performing one by a small margin (the values of the points in the graph can be found in Appendix B). In other words, usage of windows (the default value of READv1 was set to 1Mb without testing) seems to be reducing the performance of READv2.

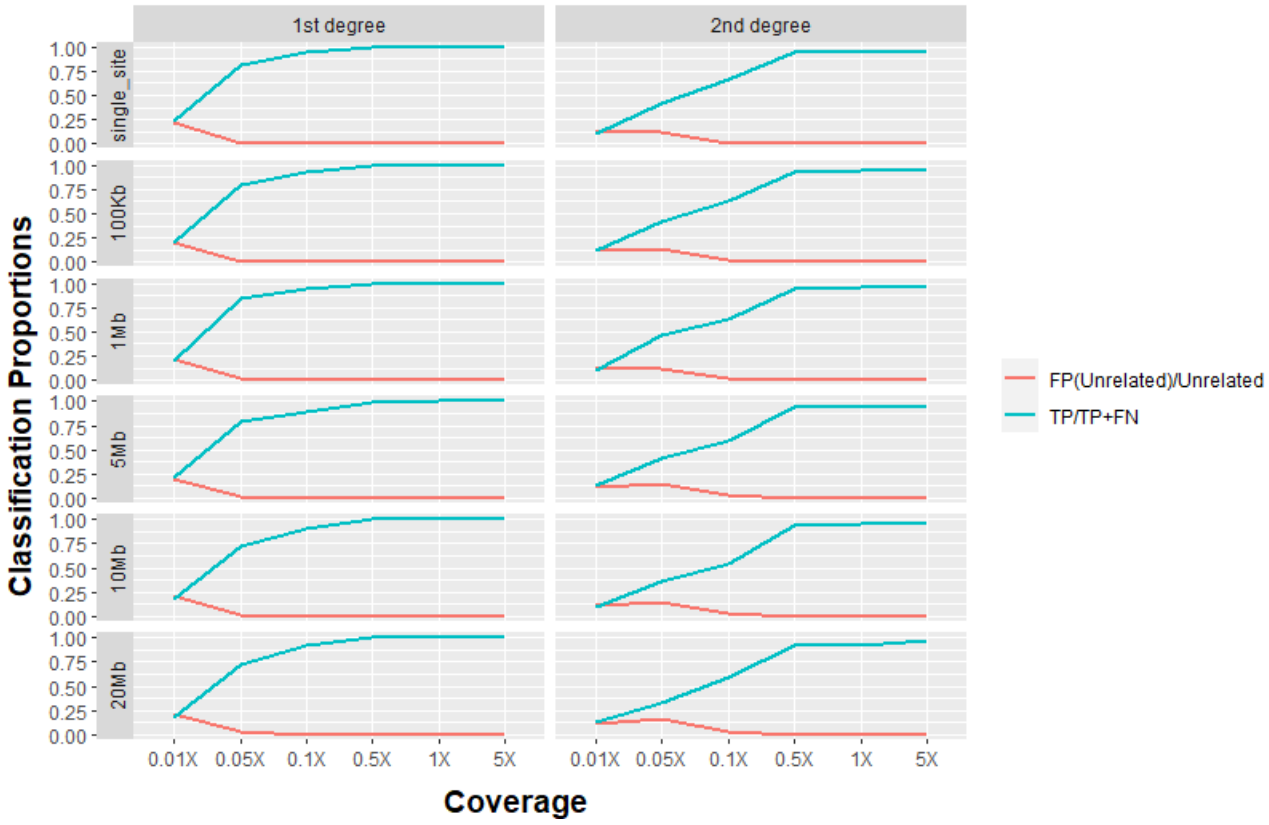


Figure 4: The sensitivity and false positive rates of READv2 for 1st degree and 2nd degree pairs. The analyses have been done for varying window sizes (100Kb, 1Mb, 5Mb, 10Mb and 20Mb) as well as for single site approaches for varying sequencing coverages (0.01X, 0.05X, 0.5X, 1X, 5X). The red line shows the proportion of false positive unrelated pairs (the pairs that are in either first degree or second degree but classified as unrelated) to all unrelated pairs. The blue line shows the sensitivity of READv2.

3.3 Variance of Pairs in First Degree Relationship

As explained in the Introduction section, the variance of non-shared alleles (P0) at the windows along the genome is expected to be higher in sibling pairs than parent-offspring pairs. In order to show this, the variance and the mean of P0 per pair was plotted for different windows and coverages (Figure 5). According to the plot, as expected, the mean normalized P0 of parent-offspring pairs and sibling pairs is the same on average but there is more variation for siblings. The difference in the variance can be only seen for high coverages, while the values for the low coverages seem to be the same. Additionally, the difference is more visible in the window size of 20Mb compared to the others and it is generally more

when the window sizes are larger. However, no clean separation of the values can be made with this plot, and the magnitude of the variance seem to be highly coverage dependent.

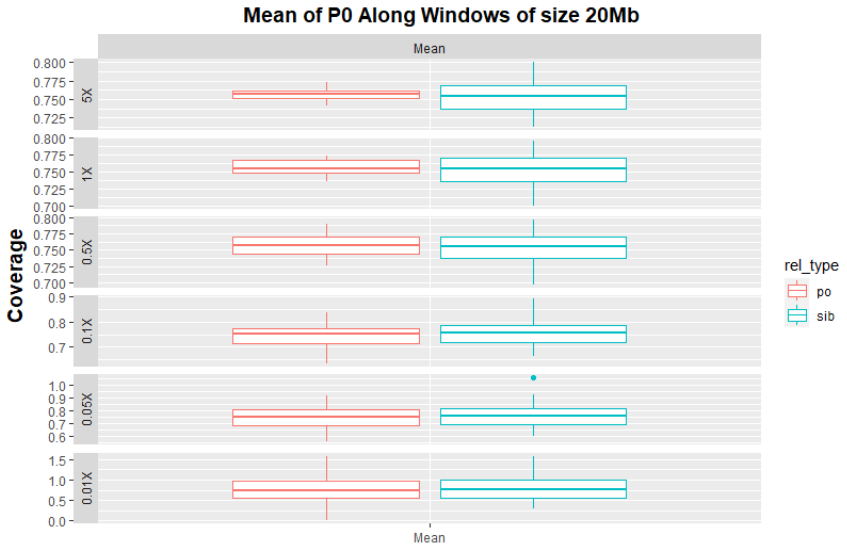
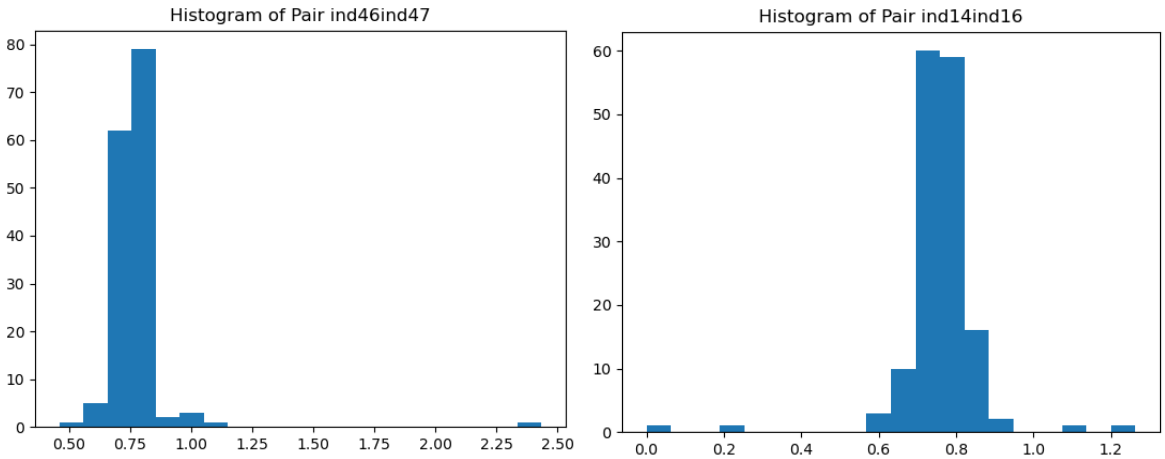


Figure 5: Mean and Variance of Normalized P0 values along the windows of varying sizes for varying coverages. The mean and variance values for window sizes of 100Kb and 20Mb are shown(1Mb and 10Mb plots can be found in Appendix C), for coverages of 5X, 1X, 0.5X, 0.1X, 0.05X, 0.01X. The blue and red bars show siblings and parent-offspring pairs, respectively.

Therefore, histogram plots of normalized P0 values of windows in 5X coverage data were created (Figure 6) for 10 random pairs in the parent-offspring and sibling relationship categories separately. The reason for this was to determine which window size yields a histogram closer to the expected result which is parent-offspring having one peak around the mean (i.e. first degree P0), and sibling pairs having three peaks for identical parts, first degree parts and unrelated parts of the genome.



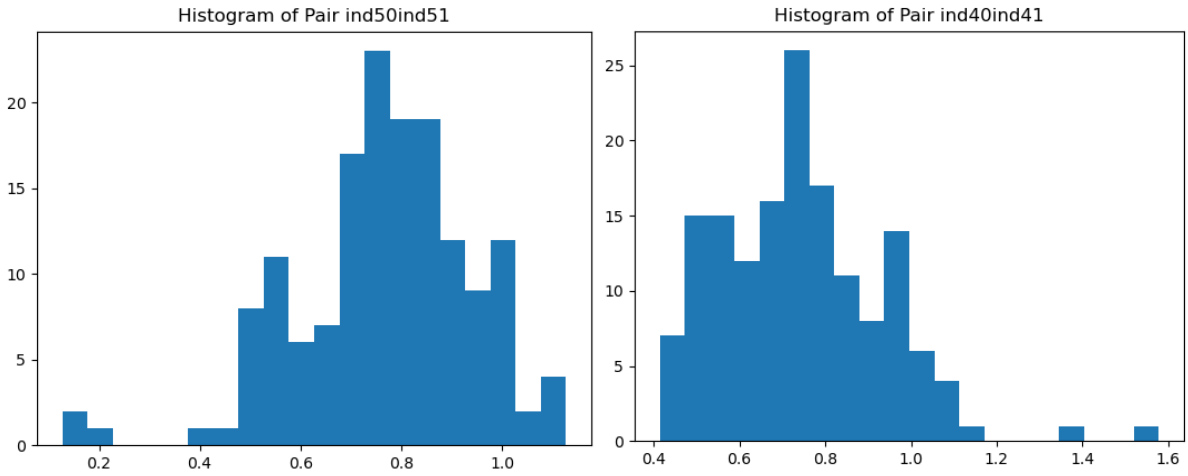


Figure 6: Example Histogram Plots of the P0 values of windows of size 20Mb along the genome with 5X coverage. Example histogram plots for parent-offspring pairs(top) and sibling pairs(bottom) are shown. 20Mb window size roughly resulted in three peaks in siblings and one peak in parent-offspring pairs as expected.

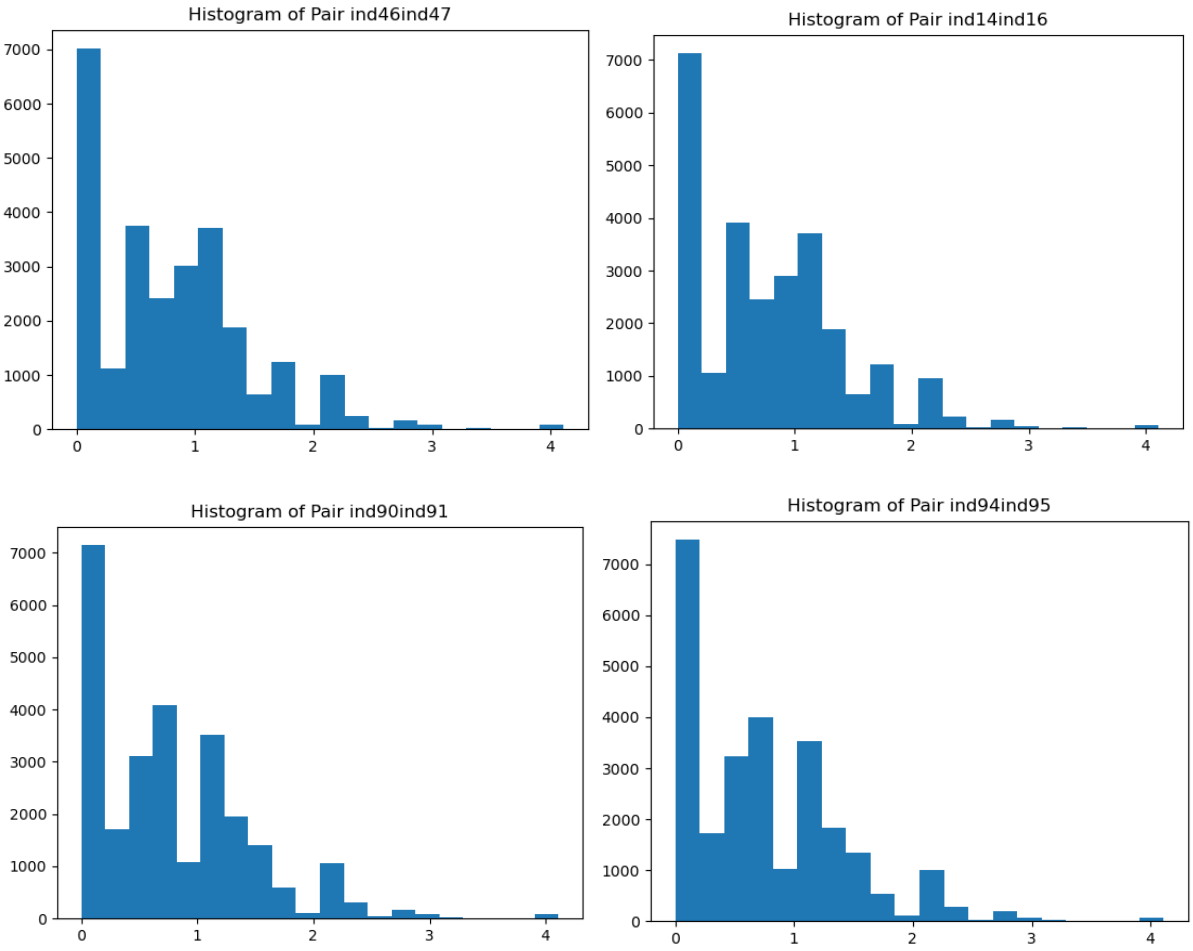


Figure 7: Example Histogram Plots of the Normalized P0 values of windows of size 100Kb along the genome with 5X coverage. Example histogram plots for parent-offspring pairs(top) and sibling pairs(bottom) are shown. 100Kb window size did not result in the expected shape of the histograms. Histogram plots for 1Mb and 10Mb can be found in Appendix D.

According to the histograms, a window size of 20Mb resulted in plots closest to the expectation (Figure 5) compared to smaller window sizes of, for instance, 100Kb (Figure 7). Therefore, a window size of 20Mb was used to calculate the proportion of all windows, in which normalized P0 values do not fall between 0.625 and 0.90625, i.e. the windows that would be classified as identical or unrelated (Figure 8). The figure gives a clear separation of the values for parent-offspring and sibling pairs for higher coverages such as 5X, 1X and 0.5X. As explained above, the expectation was that sibling pairs have more of these windows than parent-offspring pairs, and according to the plot, the cut-off values to distinguish the pairs as parent-offspring or sibling pairs was visually determined as values below 0.3 to parent-offspring, values between 0.35 and 0.6 to siblings, and values above 0.6 is not specified to any type, and they are left as unknown, as well as the values between 0.3 and 0.35. For coverages lower than 0.5X, both categories have similar and high values, as a matter of fact, the values exceed 0.6 only in lower coverages where both the parent-offspring and sibling pairs have those high values. Additionally, the interval of 0.3 and 0.35 seems to be the region where both categories sometimes overlap with high coverages.

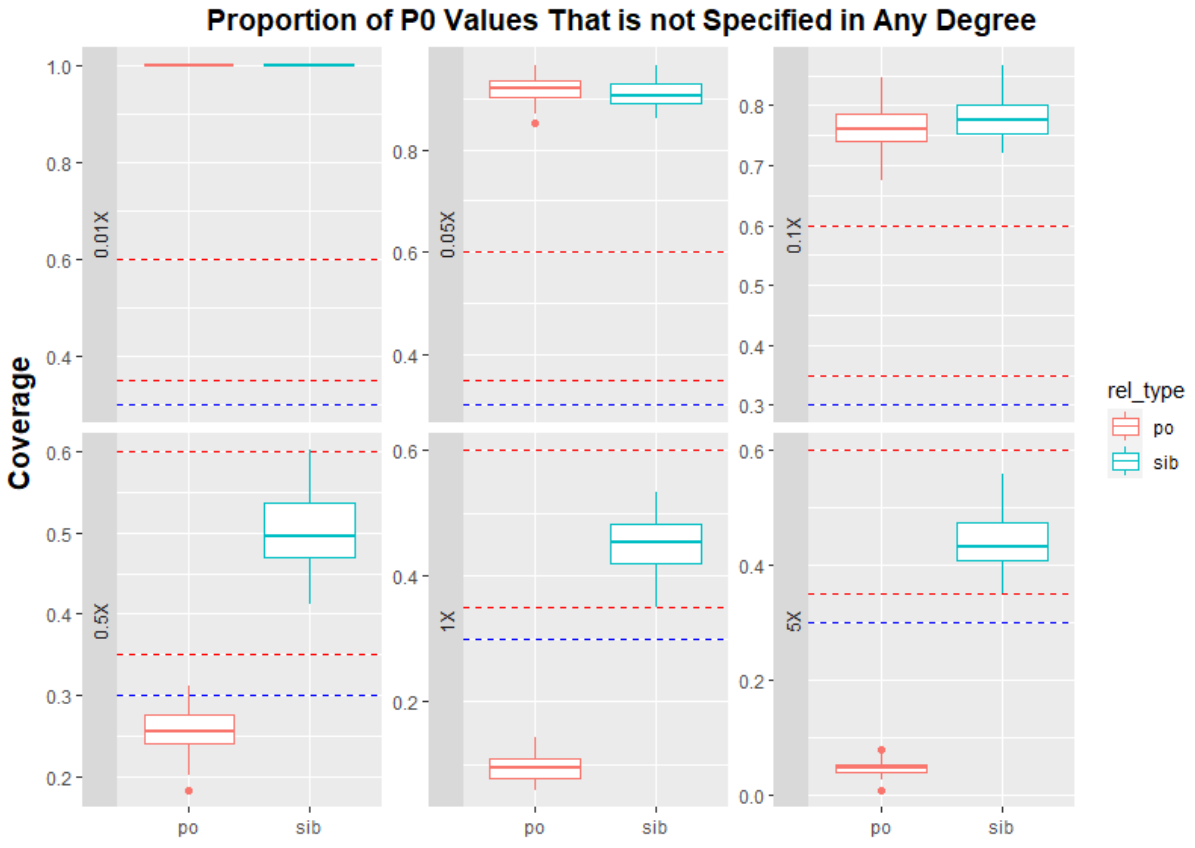


Figure 8: Proportion of Windows that cannot be specified in any degree for different coverages. The analysis was done using the window size of 20Mb. The red and blue boxes represent parent-offspring and sibling pairs, respectively. The area under blue and between red dashed lines represent the cut-off intervals for parent-offspring and sibling classifications, respectively.

3.4 Testing First Degree Classification with 1000 Genomes Data

In order to test the effectiveness of the first-degree classification method with an independent dataset, the 1000 Genomes data from CHS (Southern Han Chinese, China) and ASW (African Ancestry in Southwest USA) was used. The test was carried out with full dataset (2,458,861 SNPs), 50% (roughly 1,250,000 SNPs), 25% (roughly 625,000 SNPs), 10% (roughly 250,000 SNPs), 5% (roughly 125,000 SNPs) and 1% (roughly 25,000 SNPs) of the SNPs, to be able to determine the stress-point, i.e. the point where the performance starts to decline, of the method (Figure 9). The results show that the method is functioning well in terms of the proportion of correct matches (nearly 95% for parent-offspring pairs and 100% of the sibling pairs) and low amounts of FP (0% for both), wrong category classification (0% for both) and N/A rate (roughly 5% for parent-offspring pairs and 0% for sibling pairs), in an independent empirical dataset for number of SNPs of 125,000 and larger. However, in the case of 1% of the SNPs, it performs poorly. Nearly all the parent-offspring pairs seem to be classified as siblings and two as N/A but no TPs, and 3 of the 8 sibling pairs (37.5%) seem to have turned into N/A pairs while the rest of the samples (5 pairs) were identified correctly. AWS was excluded from the study in the end, as discrepancies between the reported degrees of relationship in the sample information data shared with the dataset were identified.

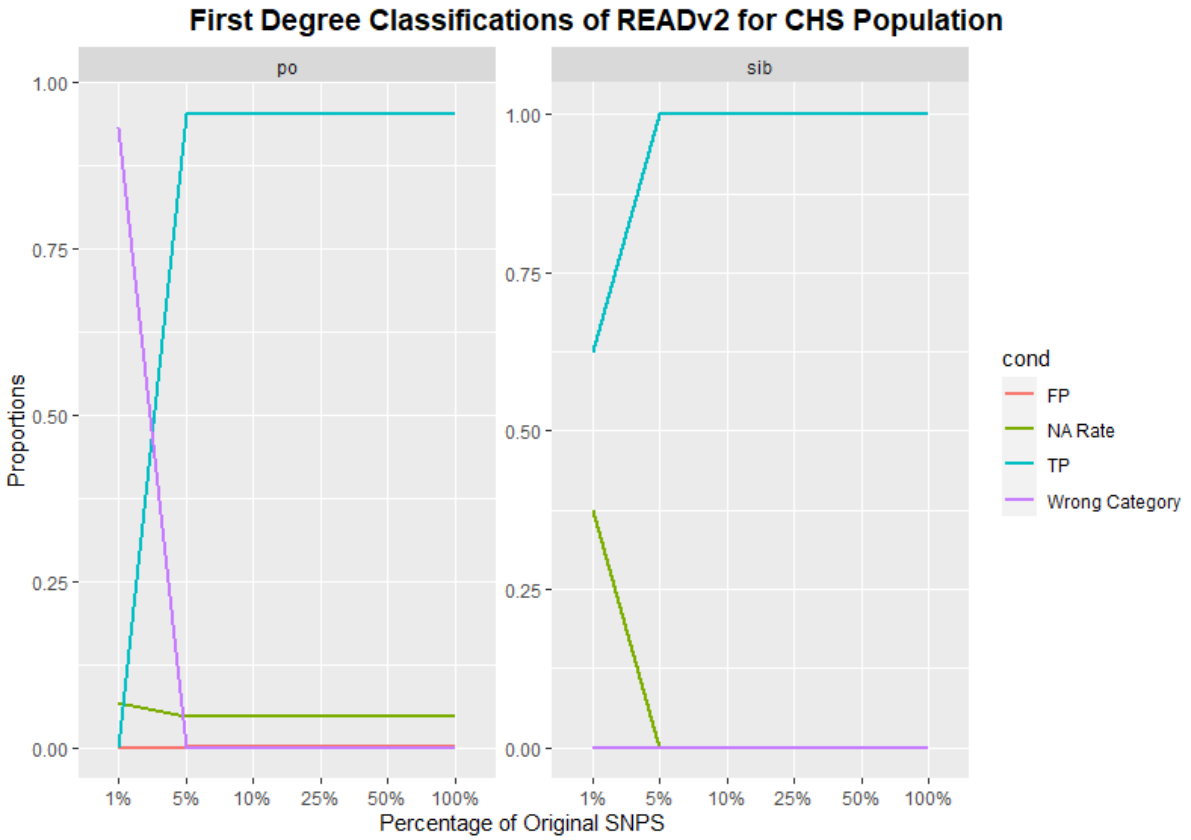


Figure 9: Test of the First-Degree Classification Method on Empirical Data from 1000 Genomes. The lines represent correctly classified parent-offspring/sibling pairs among all the parent-offspring/sibling pairs (TP), wrongly classified parent-offspring/sibling pairs among all parent-offspring/sibling pairs (Wrong Category), parent-offspring/sibling pairs classified as NA among all the parent-offspring/sibling pairs (NA Rate), and unrelated (not parent-offspring/sibling pairs) pairs classified

as parent-offspring/sibling among all unrelated pairs (FP). The TP rates seem to be high (100% for siblings, and a bit lower than 100% for the parent-offspring pairs) for percentages larger than 5%.

4 Discussion

4.1 Re-implementation of READ

One of the first goals of the project was to re-implement READ, which was originally implemented in Python2 and R (Monroy Kuhn *et al.* 2018), in Python3 in order to avoid compatibility and support issues as well as adding new features to READ such as being able to read binary .bed files which takes up less space than plain-text .tped files. The initial re-implementation took more time to run than READv1 and used much more memory in the process. This could be explained by the fact that while READv1 does not hold the data in the memory, instead it reads the input line by line and writes intermediate results in a file to be read by the R script. The new implementation holds the data in the memory and performs all operations in the same script. Furthermore, the analysis with half of the SNPs showed that the number of SNPs seems to be in a linear trend with the runtime and memory usage, while the number of individuals is in a quadratic relationship. This could be explained by the fact that as the number of individuals increases, the number of pairs increases quadratically. Therefore, it takes more resources to perform the analysis for more individuals. Additionally, the SNPs are compared linearly (i.e. once for every available locus), therefore, when the number of SNPs increases, the resource usage increase linearly.

The pairwise comparison was then implemented with NumPy array comparisons instead of pairwise comparisons in loops, in order to increase the efficiency of the script. NumPy arrays efficiently store and access the data, and many NumPy operations were implemented in C programming language, avoiding the high cost of loops in Python (Harris *et al.* 2020). Because of that, the expectation was to decrease both the runtime and memory cost compared to the initial READv2 implementation, and the results of the analysis of runtime and memory usage of READv2 have confirmed the expectations. Additionally, the relationship of SNP size and sample size with runtime and memory usage seem to hold in this case as well. Although the memory usage of READv2 have improved with the usage of NumPy arrays, it is still performing poorly compared to READv1, due to the reasons explained above.

4.2 Window size and Single-Site Approach Comparison

Another goal of the project was to compare the effects of window-based or single-site approaches which were not formally compared before. The window-based approach that is the default mode for READv1 performed poorly compared to a single-site approach, in terms of runtime and memory usage (Figure 3). In fact, the single-site approach of READv2 seems to perform similar to window-based approach with half of the SNPs in terms of runtime, and to

window-based approach with half of the individuals in terms of memory usage. Since the single-site approach of READv2 was implemented in a way that uses the whole sequence as one window, the number of generated results per pair is just one line, instead of one line per each window evaluated. Therefore, memory and runtime costs were expected to be lower for the single-site approach than the window-based approach for READv2. Additionally, in the way that READv2 is implemented right now, the genome data is first divided into windows, and then each pair is compared for each window. This, and having only one line of result for each pair might explain the similarities to the case of half of the SNPs in runtime and to the case of half of the individuals in memory usage, respectively.

The analysis with varying window sizes and single site across different coverage situations using the simulated data with known relationships also resulted in favour of the single-site approach for low amounts of data. The power of the method (i.e. sensitivity) for different window sizes and single-site was similar for sufficient amount of data (i.e. the data with higher coverages, in this case 5X, 1X and 0.5X for first-degree classification; 5X and 1X for second-degree classifications). The poorer performance of 2nd degree classifications for lower coverages suggests that the performance of READv2 is more susceptible to missingness of data for second degree relationships, which was observed in READv1 as well (Monroy Kuhn *et al.* 2018). The reason for this is simply because the interval for 2nd degree relationships is narrower, therefore, the noise tends to push some pairs into "unrelated" category. For the lower coverages, the single-site approach of READv2 seems to be the best performing one by a small margin (Appendix B), and the performances seem to worsen as the window sizes increases. However, the single-site approach could not have been used in the successive analyses of the project, because, in order to evaluate the variance of the windows, normalized P0 values from multiple windows were needed, as explained above, single-site approach yields one normalized P0 value for the entire genome. Another disadvantage of the single-site approach arising from this is that the confidence for classification cannot be estimated, therefore, the uncertainty of the results cannot be assessed. If this was implemented in a way that estimates uncertainty, the single site approach might have gotten slower as well.

4.3 Variance of Pairs in First Degree Relationship

As expected, the mean P0 values for different window sizes were the same for parent-offspring and sibling pairs, while the variance of windows across the genome seems to differ. However, there is no consistent quantitative difference between the variance levels of parent-offspring and siblings that could be used as a cut-off value to classify them separately in each coverage or window size. In addition to this, all variances seem to increase with lower coverages.

Parent-offspring and sibling pairs were also expected to separate in terms of distribution of normalized P0 values of windows, since an offspring receives exactly one chromosome directly from a parent, the difference between a parent and an offspring is expected to be the

same along the genome. However, siblings can inherit different chromosomes from their parents, as well as the same ones. Therefore, the difference between two sibling genomes is expected to differ more along the genome: some parts identical, some parts in first-degree pattern, and some parts in unrelated pattern. As a result of this, the normalized P0 histogram of a parent-offspring pair is expected to have one peak near the mean that is in first degree classification interval, while the same histogram that belongs to a sibling pair is expected to have three peaks located near a value in the identical/twin interval, a value near the mean, and a value in the unrelated interval. In other words, a sibling pair is expected to contain more windows that cannot be classified as in first- or second-degree relationship, compared to a parent-offspring pair.

The histograms showed that a window size of 20Mb yields the results closest to the expectations. The other window sizes, such as window size of 100Kb, resulted in peaks close to 0, which are windows of almost no difference. This might be because of the fact that smaller windows mean less SNPs, hence, the P0 value per window is noisier, and therefore, the classification of individual windows is more difficult. Consequently, a window size of 20Mb was used to determine the cut-off values in first degree relationship classification.

The analysis of proportion of windows that cannot be specified as first or second degree showed clear separation of parent-offspring and sibling pairs in coverages higher than and equal to 0.5X. Based on the clear separations, the cut-offs for parent-offspring (smaller than 0.3), sibling (between 0.35 and 0.6) and N/A (other values) categories were defined. The inclusion of N/A category was in order to allow reducing wrong classifications (e.g. parent-offspring pairs as sibling pairs) in the case of low amounts of data. The proportions for lower coverages (0.1X, 0.05X, 0.01X) seem to converge to 1 regardless of the type of relationship and fall out of the classification intervals for siblings and parent-offspring mentioned above, so they would be classified as N/A. Similarly, this can be seen in the histograms where the variation goes up to 4 which is approximately the maximum normalized P0 value for the simulated data (maximum P0 value of 1 divided by the approximate median of 0.25) which means that there are mostly 0s and 4s, and very little in the middle where the classification takes place.

The same trend can be seen in the case of the empirical data from the CHS population from the 1000 Genomes Project as well. The test shows that the majority of parent-offspring pairs were classified as siblings, and a large part of the siblings classified as N/A (as in the specific type cannot be determined) when using 1% of the SNPs (roughly 25,000 SNPs). This could be a result of the abovementioned process: as the coverage (or the amount of data) decreases for individuals, P0 values per window are spread more widely and the proportion of windows that cannot be specified as first- or second-degree increases. Therefore, the results for parent-offspring pairs are now overlapping the sibling classification interval. Similarly, the true sibling pairs are now pushed into to the N/A interval.

Additionally, the test seems to be working when using 5% (roughly 125,000 SNPs) or more of all the SNPs and it also worked for 0.5X coverage (roughly 50,000 overlapping SNPs) in the simulation data. However, it seems to fail for 1% (25,000 SNPs) and all the classifications below 0.1X (roughly 2,000 overlapping SNPs) for simulations were N/A. Therefore, the interval where the method fails might have been hit, unintentionally. Unfortunately, the range between 0.1X and 0.5X is where many published ancient individuals fall (Mallick *et al.* 2023) making it crucial to avoid false classifications.

Although the abovementioned test presents many results in terms of usability of this method for an independent dataset, it does not show the stress-point of the method (i.e. where the TP values starts to drop), as well as the point where all the pairs have been pushed up to the N/A zone.

4.4 Future Directions

Resource usage: Although READv2 works faster than READv1, it can be further improved by first performing the pairwise comparisons for all sites (i.e. the same way in single-site approach) and then dividing the resulting NumPy array into windows, instead of dividing the genome into windows and doing the pairwise comparisons for each window, as explained above. This way, the number of operations performed might decrease, and this might improve the runtime, and possibly the memory costs.

Distinguishing parent offspring/sibling pairs: As explained above, the test with the 1000 Genomes population showed that the first-degree classification method can be used with an independent dataset, and it performs as expected for larger subsets of the full data. It also showed that, when using 1% of the SNPs, most of the parent-offspring pairs have been predicted to be siblings, which might be misleading in a real scenario. One way to solve this could be implement a way to express the confidence levels for the classification. This way, the reliability of the results can be shown to the user.

Besides the results that can be observed from the test, there are some conclusions that cannot be drawn, such as at what minimum threshold would all the pairs be pushed into the N/A zone, or when would the TP values start to drop (the stress-point of the method). These could be determined by further analysis including e.g., 0.5%, 2%, 3% and 4% of the SNPs. Furthermore, including more sibling pairs (i.e. including data from more populations) might decrease the effect of noise and provide more resolution on the exact performances for sibling predictions.

Release as a software: Since READv2 is a method intended to be used in active aDNA research, it needs to be made more user friendly such that any researcher can run READv2 with desired parameters and without needing to change it manually in the script. Therefore, the two different approaches, window-based and single-site, need to be set as two different options that can be set with different flags. However, if the single-site approach is used, there

should be a way to express the uncertainties, or a way to be able to use the first-degree classification method. At least, the latter can be done, possibly, going over the data twice: once to determine the degrees with the single-site approach, and second time to distinguish the pairs in a first-degree relationship with the window-based approach. However, the effect of this on runtime and memory usage needs to be investigated as well. Another consideration should be on changing the default window size from 1Mb to a smaller one. Similarly, the effect on the required resources, as well as the performance of READv2 with independent datasets could give an idea on how to make this decision.

5 Ethical concerns and conflict of interest

The project does not require any permits nor include any ethical concerns.

6 Acknowledgements

I would like to thank my supervisor Torsten Günther and my subject reader Thijessen Naidoo for their guidance and mentorship, as well as past and current members of Günther Lab and Human Evolution Programme for their ideas, inputs, and challenging questions in the presentations throughout my project. Without the help of the members of Somel Lab from METU, I would not be able to make sense of the data and how to use it in ages. A special thanks to my dear girlfriend Yasemin for her contribution to the aesthetics of the plots, without her comments and ideas, none of the plots would be pleasing to examine. Last but not least, without the support of my family for my master's education, especially without the sponsorship of my sister, Simge, I could not have done any of these mentioned works.

References

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu Y, Chang Y, Feng Q, Fang X, Guo X, Jian M, Jiang H, Jin X, Lan T, Li G, Li J, Li Y, Liu S, Liu X, Lu Y, Ma X, Tang M, Wang B, Wang G, Wu H, Wu R, Xu X, Yin Y, Zhang D, Zhang W, Zhao J, Zhao M, Zheng X, Gupta N, Gharani N, Toji LH, Gerry NP, Resch AM, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Grocock R, Humphray S, James T, Kingsbury Z, Sudbrak R, Albrecht MW, Amstislavskiy VS, Borodina TA, Lienhard M, Mertes F, Sultan M, Timmermann B, Yaspo ML, Fulton L, Ananiev V, Belaia Z, Beloslyudtsev D, Bouk N, Chen C, Church D, Cohen R, Cook C, Garner J, Hefferon T, Kimelman M, Liu C, Lopez J, Meric P, O'Sullivan C, Ostapchuk Y, Phan L, Ponomarov S, Schneider V, Shekhtman E, Sirotkin K, Slotta D, Zhang H, Balasubramaniam S, Burton J, Danecek P, Keane TM, Kolb-Kokocinski A, McCarthy S, Stalker J, Quail M, Davies CJ, Gollub J, Webster T, Wong B, Zhan Y, Campbell CL, Kong Y, Marcketta A, Yu F, Antunes L, Bainbridge M, Sabo A, Huang Z, Coin LJM, Fang L, Li Q, Li Z, Lin H, Liu B, Luo R, Shao H, Xie Y, Ye C, Yu C, Zhang F, Zheng H, Zhu H, Alkan C, Dal E, Kahveci F, Garrison EP, Kural D, Lee WP, Leong WF, Stromberg M, Ward AN, Wu J, Zhang M, Daly MJ, DePristo MA, Handsaker RE, Banks E, Bhatia G, Del Angel G, Genovese G, Li H, Kashin S, McCarroll SA, Nemesh JC, Poplin RE, Yoon SC, Lihm J, Makarov V, Gottipati S, Keinan A, Rodriguez-Flores JL, Rausch T, Fritz MH, Stütz AM, Beal K, Datta A, Herrero J, Ritchie GRS, Zerbino D, Sabeti PC, Shlyakhter I, Schaffner SF, Vitti J, Cooper DN, Ball E V., Stenson PD, Barnes B, Bauer M, Cheetham RK, Cox A, Eberle M, Kahn S, Murray L, Peden J, Shaw R, Kenny EE, Batzer MA, Konkel MK, Walker JA, MacArthur DG, Lek M, Herwig R, Ding L, Koboldt DC, Larson D, Ye K, Gravel S, Swaroop A, Chew E, Lappalainen T, Erlich Y, Gymrek M, Willems TF, Simpson JT, Shriver MD, Rosenfeld JA, Bustamante CD, Montgomery SB, De La Vega FM, Byrnes JK, Carroll AW, DeGorter MK, Lacroute P, Maples BK, Martin AR, Moreno-Estrada A, Shringarpure SS, Zakharia F, Halperin E, Baran Y, Cerveira E, Hwang J, Malhotra A, Plewczynski D, Radew K, Romanovitch M, Zhang C, Hyland FCL, Craig DW, Christoforides A, Homer N, Izatt T, Kurdoglu AA, Sinari SA, Squire K, Xiao C, Sebat J, Antaki D, Gujral M, Noor A, Ye K, Burchard EG, Hernandez RD, Gignoux CR, Haussler D, Katzman SJ, Kent WJ, Howie B, Ruiz-Linares A, Dermitzakis ET, Devine SE, Kang HM, Kidd JM, Blackwell T, Caron S, Chen W, Emery S, Fritsche L, Fuchsberger C, Jun G, Li B, Lyons R, Scheller C, Sidore C, Song S, Sliwerska E, Taliun D, Tan A, Welch R, Wing MK, Zhan X, Awadalla P, Hodgkinson A, Li Y, Shi X, Quitadamo A, Lunter G, Marchini JL, Myers S, Churchhouse C, Delaneau O, Gupta-

Hinch A, Kretzschmar W, Iqbal Z, Mathieson I, Menelaou A, Rimmer A, Xifara DK, Oleksyk TK, Fu Y, Liu X, Xiong M, Jorde L, Witherspoon D, Xing J, Browning BL, Browning SR, Hormozdiari F, Sudmant PH, Khurana E, Tyler-Smith C, Albers CA, Ayub Q, Chen Y, Colonna V, Jostins L, Walter K, Xue Y, Gerstein MB, Abyzov A, Balasubramanian S, Chen J, Clarke D, Fu Y, Harmanci AO, Jin M, Lee D, Liu J, Mu XJ, Zhang J, Zhang Y, Hartl C, Shakir K, Degenhardt J, Meiers S, Raeder B, Casale FP, Stegle O, Lameijer EW, Hall I, Bafna V, Michaelson J, Gardner EJ, Mills RE, Dayama G, Chen K, Fan X, Chong Z, Chen T, Chaisson MJ, Huddleston J, Malig M, Nelson BJ, Parrish NF, Blackburne B, Lindsay SJ, Ning Z, Zhang Y, Lam H, Sisu C, Challis D, Evani US, Lu J, Nagaswamy U, Yu J, Li W, Habegger L, Yu H, Cunningham F, Dunham I, Lage K, Jespersen JB, Horn H, Kim D, Desalle R, Narechania A, Sayres MAW, Mendez FL, Poznik GD, Underhill PA, Mittelman D, Banerjee R, Cerezo M, Fitzgerald TW, Louzada S, Massaia A, Yang F, Kalra D, Hale W, Dan X, Barnes KC, Beiswanger C, Cai H, Cao H, Henn B, Jones D, Kaye JS, Kent A, Kerasidou A, Mathias R, Ossorio PN, Parker M, Rotimi CN, Royal CD, Sandoval K, Su Y, Tian Z, Tishkoff S, Via M, Wang Y, Yang H, Yang L, Zhu J, Bodmer W, Bedoya G, Cai Z, Gao Y, Chu J, Peltonen L, Garcia-Montero A, Orfao A, Dutil J, Martinez-Cruzado JC, Mathias RA, Hennis A, Watson H, McKenzie C, Qadri F, LaRocque R, Deng X, Asogun D, Folarin O, Happi C, Omoniwa O, Stremlau M, Tariyal R, Jallow M, Joof FS, Corrah T, Rockett K, Kwiatkowski D, Kooner J, Hien TT, Dunstan SJ, ThuyHang N, Fonnie R, Garry R, Kanneh L, Moses L, Schieffelin J, Grant DS, Gallo C, Poletti G, Saleheen D, Rasheed A, Brooks LD, Felsenfeld AL, McEwen JE, Vaydylevich Y, Duncanson A, Dunn M, Schloss JA. 2015. A global reference for human genetic variation. Nature 2015 526:7571 526: 68–74.

Barlow A, Hartmann S, Gonzalez J, Hofreiter M, Paijmans JLA. 2020. Consensify: A Method for Generating Pseudohaploid Genome Sequences from Palaeogenomic Datasets with Reduced Error Rates. Genes, doi 10.3390/GENES11010050.

Brace S, Diekmann Y, Booth TJ, van Dorp L, Faltyskova Z, Rohland N, Mallick S, Olalde I, Ferry M, Michel M, Oppenheimer J, Broomandkhoshbacht N, Stewardson K, Martiniano R, Walsh S, Kayser M, Charlton S, Hellenthal G, Armit I, Schulting R, Craig OE, Sheridan A, Parker Pearson M, Stringer C, Reich D, Thomas MG, Barnes I. 2019. Ancient genomes indicate population replacement in Early Neolithic Britain. Nature Ecology & Evolution 2019 3:5 3: 765–771.

Caballero M, Seidman DN, Qiao Y, Sannerud J, Dyer TD, Lehman DM, Curran JE, Duggirala R, Blangero J, Carmi S, Williams AL. 2019. Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. PLOS Genetics 15: e1007979.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. GigaScience 4: 7.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM. 2021. Twelve years of SAMtools and BCFtools. GigaScience, doi 10.1093/GIGASCIENCE/GIAB008.

Eddy SR. 1996. Hidden Markov models. Current Opinion in Structural Biology 6: 361–365.

Fernandes D, Sirak K, Novak M, Finarelli JA, Byrne J, Connolly E, Carlsson JEL, Ferretti E, Pinhasi R, Carlsson J. 2017. The Identification of a 1916 Irish Rebel: New Approach for Estimating Relatedness From Low Coverage Homozygous Genomes. Scientific Reports 7: 41529.

Fernandes DM, Cheronet O, Gelabert P, Pinhasi R. 2021. TKGWV2: an ancient DNA relatedness pipeline for ultra-low coverage whole genome shotgun data. Scientific Reports 2021 11:1 11: 1–9.

Fowler C, Olalde I, Cummings V, Armit I, Büster L, Cuthbert S, Rohland N, Cheronet O, Pinhasi R, Reich D. 2022. A high-resolution picture of kinship practices in an Early Neolithic tomb. 584 | Nature |, doi 10.1038/s41586-021-04241-4.

Frånberg M. 2021. mfranberg/libplinkio: A small C and Python library for reading PLINK genotype files. online 21 February 2021: https://github.com/mfranberg/libplinkio. Accessed 5 May 2023.

Hanghøj K, Moltke I, Andersen PA, Manica A, Korneliussen TS. 2019. Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. GigaScience 8: 1–9.

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE. 2020. Array programming with NumPy. Nature 2020 585:7825 585: 357–362.

Henden L, Lee S, Mueller I, Barry A, Bahlo M. 2018. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. PLoS Genetics, doi 10.1371/JOURNAL.PGEN.1007279.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25: 1754–1760.

Lipatov M, Sanjeev K, Patro R, Veeramah KR. 2015. Maximum Likelihood Estimation of Biological Relatedness from Low Coverage Sequencing Data. doi 10.1101/023374.

Mallick S, Micco A, Mah M, Ringbauer H, Lazaridis I, Olalde I, Patterson N, Reich D. 2023. The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes. bioRxiv 2023.04.06.535797.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics 26: 2867–2873.

Marcus JH, Posth C, Ringbauer H, Lai L, Skeates R, Sidore C, Beckett J, Furtwängler A, Olivieri A, Chiang CWK, Al-Asadi H, Dey K, Joseph TA, Liu CC, Der Sarkissian C, Radzevičiūtė R, Michel M, Gradoli MG, Marongiu P, Rubino S, Mazzarello V, Rovina D, La Fragola A, Serra RM, Bandiera P, Bianucci R, Pompianu E, Murgia C, Guirguis M, Orquin RP, Tuross N, van Dommelen P, Haak W, Reich D, Schlessinger D, Cucca F, Krause J, Novembre J. 2020. Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia. Nature Communications 2020 11:1 11: 1–14.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, Sirak K, Gamba C, Jones ER, Llamas B, Dryomov S, Pickrell J, Arsuaga JL, De Castro JMB, Carbonell E, Gerritsen F, Khokhlov A, Kuznetsov P, Lozano M, Meller H, Mochalov O, Moiseyev V, Guerra MAR, Roodenberg J, Vergès JM, Krause J, Cooper A, Alt KW, Brown D, Anthony D, Lalueza-Fox C, Haak W, Pinhasi R, Reich D. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. Nature 2015 528:7583 528: 499–503.

McKinney W. 2010. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, doi 10.25080/majora-92bf1922-00a.

Monroy Kuhn JM, Jakobsson M, Günther T. 2018. Estimating genetic kin relationships in prehistoric populations. doi 10.1371/journal.pone.0195491.

Nyerki E, Kalmár T, Schütz O, Lima RM, Neparáczki E, Török T, Maróti Z. 2023. correctKin: an optimized method to infer relatedness up to the 4th degree from low-coverage ancient human genomes. Genome Biology 24: 1–21.

Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, Rohland N, Mallick S, Szécsényi-Nagy A, Mittnik A, Altena E, Lipson M, Lazaridis I, Harper TK, Patterson N, Broomandkhoshbacht N, Diekmann Y, Faltyskova Z, Fernandes D, Ferry M, Harney E, De Knijff P, Michel M, Oppenheimer J, Stewardson K, Barclay A, Alt KW, Liesau C, Rios P, Blasco C, Miguel JV, Garcia RM, Fernandez AA, Banffy E, Bernabo-Brea M, Billoin D, Bonsall C, Bonsall L, Allen T, Buster L, Carver S, Navarro LC, Craig OE, Cook GT, Cunliffe B, Denaire A, Dinwiddy KE, Dodwell N, Ernee M, Evans C,

Kucharik M, Farre JF, Fowler C, Gazenbeek M, Pena RG, Haber-Uriarte M, Haduch E, Hey G, Jowett N, Knowles T, Massy K, Pfrengle S, Lefranc P, Lemercier O, Lefebvre A, Martinez CH, Olmo VG, Ramirez AB, Maurandi JL, Majo T, McKinley JI, McSweeney K, Mende BG, Mod A, Kulcsar G, Kiss V, Czene A, Patay R, Endrodi A, Kohler K, Hajdu T, Szeniczey T, Dani J, Bernert Z, Hoole M, Cheronet O, Keating D, Veleminsky P, Dobe M, Candilio F, Brown F, Fernandez RF, Herrero-Corral AM, Tusa S, Carnieri E, Lentini L, Valenti A, Zanini A, Waddington C, Delibes G, Guerra-Doce E, Neil B, Brittain M, Luke M, Mortimer R, Desideri J, Besse M, Brucken G, Furmanek M, Hauszko A, Mackiewicz M, Rapinski A, Leach S, Soriano I, Lillios KT, Cardoso JL, Pearson MP, Wodarczak P, Price TD, Prieto P, Rey PJ, Risch R, Guerra MAR, Schmitt A, Serralongue J, Silva AM, Smrcka V, Vergnaud L, Zilhao J, Caramelli D, Higham T, Thomas MG, Kennett DJ, Fokkens H, Heyd V, Sheridan A, Sjogren KG, Stockhammer PW, Krause J, Pinhasi R, Haak W, Barnes I, Lalueza-Fox C, Reich D. 2018. The Beaker phenomenon and the genomic transformation of northwest Europe. Nature 2018 555:7695 555: 190–196.

Oliehoek PA, Windig JJ, van Arendonk JAM, Bijma P. 2006. Estimating Relatedness Between Individuals in General Populations With a Focus on Their Use in Conservation Programs. Genetics 173: 483–496.

Popli D, Peyrégne S, Peter BM. 2023. KIN: a method to infer relatedness from low-coverage ancient DNA. Genome biology 24: 10.

Queller DC, Goodnight KF. 1989. ESTIMATING RELATEDNESS USING GENETIC MARKERS. Evolution 43: 258–275.

Renaud G, Hanghøj K, Willerslev E, Orlando L. 2017. gargammel: a sequence simulator for ancient DNA. Bioinformatics (Oxford, England) 33: 577–579.

Ringbauer H, Huang Y, Akbari A, Mallick S, Patterson N, Reich D. 2023. ancIBD - Screening for identity by descent segments in human ancient DNA. bioRxiv 2023.03.08.531671.

Skov L, Peyrégne S, Popli D, Iasi LNM, Devièse T, Slon V, Zavala EI, Hajdinjak M, Sümer AP, Grote S, Bossoms Mesa A, López Herráez D, Nickel B, Nagel S, Richter J, Essel E, Gansauge M, Schmidt A, Korlević P, Comeskey D, Derevianko AP, Kharevich A, Markin S V., Talamo S, Douka K, Krajcarz MT, Roberts RG, Higham T, Viola B, Krivoshapkin AI, Kolobova KA, Kelso J, Meyer M, Pääbo S, Peter BM. 2022. Genetic insights into the social organization of Neanderthals. Nature 610: 519–525.

Appendix A – Python Script for Pseudo-Haploid Alleles

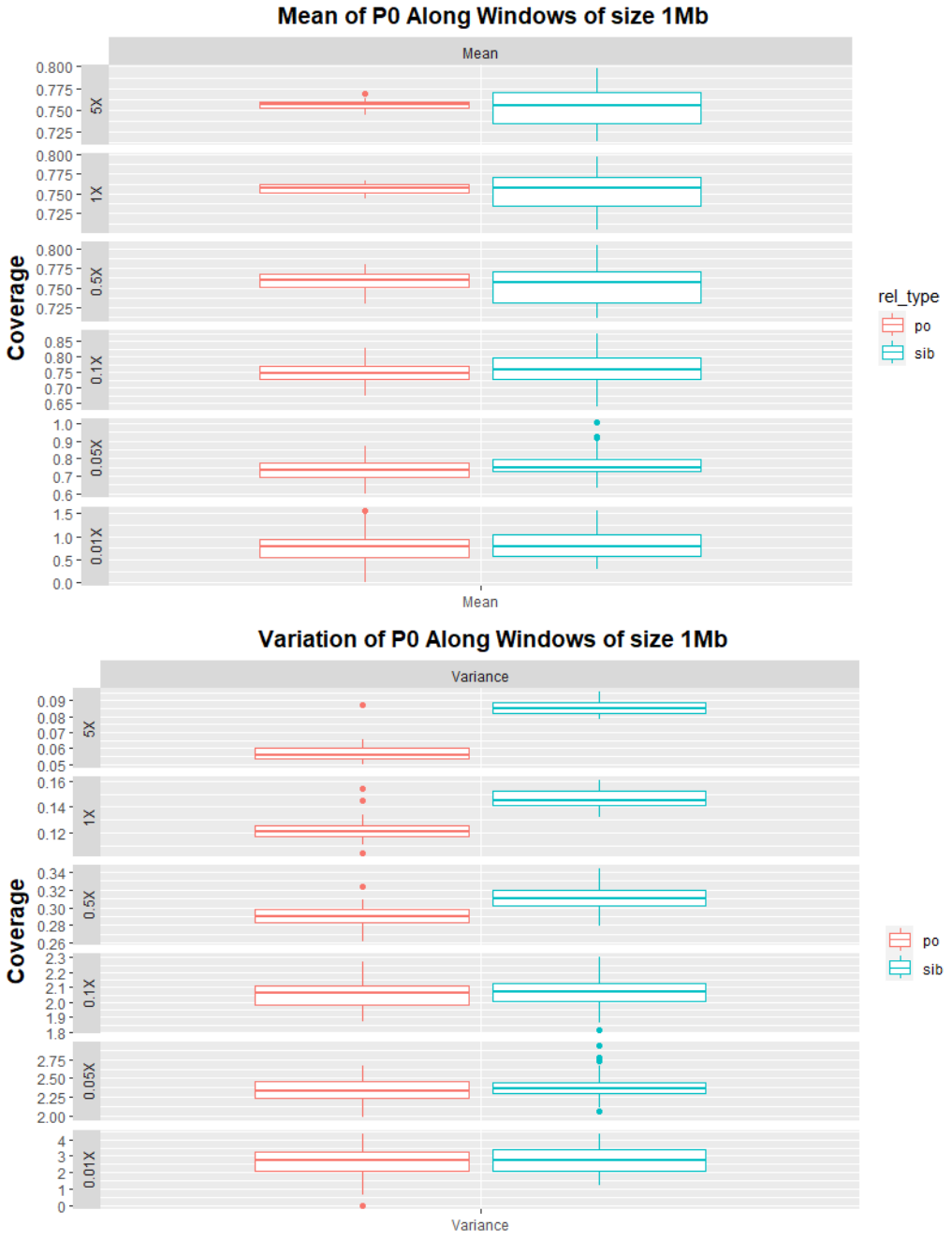
```
#!/usr/bin/env python
from numpy.random import randint
import sys
##read tped from stdin and make all sites homozygous
for l in sys.stdin:
    split=l.rstrip().split()
    for i in xrange(4,len(split),2):
        a=split[i+randint(0,2)]
        split[i]=a
        split[i+1]=a
    print ' '.join(split)
```

Appendix B – Graph Values

Degree	Values	single_site	100Kb	1Mb	10Mb	5Mb	20Mb	covs
1st Deg	Sensitivity	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	5X
1st Deg	FP	0.0010586	0.0010587	0.0010586	0.0010587	0.0010586	0.0010588	5X
2nd Deg	Sensitivity	0.9513889	0.9513889	0.9513889	0.9583333	0.9513889	0.9722222	5X
2nd Deg	FP	0.0017785	0.0018210	0.0017785	0.0017786	0.0017361	0.0017788	5X
1st Deg	Sensitivity	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1X
1st Deg	FP	0.0010586	0.0010585	0.0010586	0.0010586	0.0010585	0.0010584	1X
2nd Deg	Sensitivity	0.9513889	0.9444444	0.9513889	0.9513889	0.9444444	0.9166667	1X
2nd Deg	FP	0.0017361	0.0016512	0.0017785	0.0017361	0.0016936	0.0017357	1X
1st Deg	Sensitivity	1.0000000	1.0000000	1.0000000	1.0000000	0.9895833	1.0000000	0.5X
1st Deg	FP	0.0010586	0.0010585	0.0010586	0.0010585	0.0010586	0.0010583	0.5X
2nd Deg	Sensitivity	0.9444444	0.9305556	0.9375000	0.9444444	0.9444444	0.9236111	0.5X
2nd Deg	FP	0.0018209	0.0017783	0.0018632	0.0016936	0.0017360	0.0016510	0.5X
1st Deg	Sensitivity	0.9479167	0.9270833	0.9375000	0.9062500	0.8854167	0.9270833	0.1X
1st Deg	FP	0.0010668	0.0010705	0.0011161	0.0010364	0.0011254	0.0010802	0.1X
2nd Deg	Sensitivity	0.6666667	0.6250000	0.6250000	0.5486111	0.5972222	0.5833333	0.1X
2nd Deg	FP	0.0110096	0.0146442	0.0169564	0.0235360	0.0254956	0.0238939	0.1X
1st Deg	Sensitivity	0.8125000	0.7916667	0.8437500	0.7291667	0.7916667	0.7187500	0.05X
1st Deg	FP	0.0086878	0.0107532	0.0111807	0.0206267	0.0152047	0.0235036	0.05X
2nd Deg	Sensitivity	0.4236111	0.4166667	0.4722222	0.3611111	0.4097222	0.3333333	0.05X
2nd Deg	FP	0.1177839	0.1284144	0.1224139	0.1507525	0.1411306	0.1586993	0.05X
1st Deg	Sensitivity	0.2395833	0.2083333	0.1979167	0.1770833	0.2083333	0.1770833	0.01X
1st Deg	FP	0.2180556	0.2026100	0.2091350	0.2054186	0.2053636	0.2075115	0.01X
2nd Deg	Sensitivity	0.1111111	0.1250000	0.0972222	0.0972222	0.1250000	0.1250000	0.01X
2nd Deg	FP	0.1160240	0.1154612	0.1204618	0.1181435	0.1209592	0.1169509	0.01X

Table S1: The graph values of the Figure 4. It can be seen that the sensitivity values are equal to 1 for each window size for 1st degree individuals in high coverages (0.5X, 1X, 5X). The differences between the windows start to arise for the low coverage, and the single site approach seems to be superior to all window sizes.

Appendix C – Mean and Variance for Extra Windows



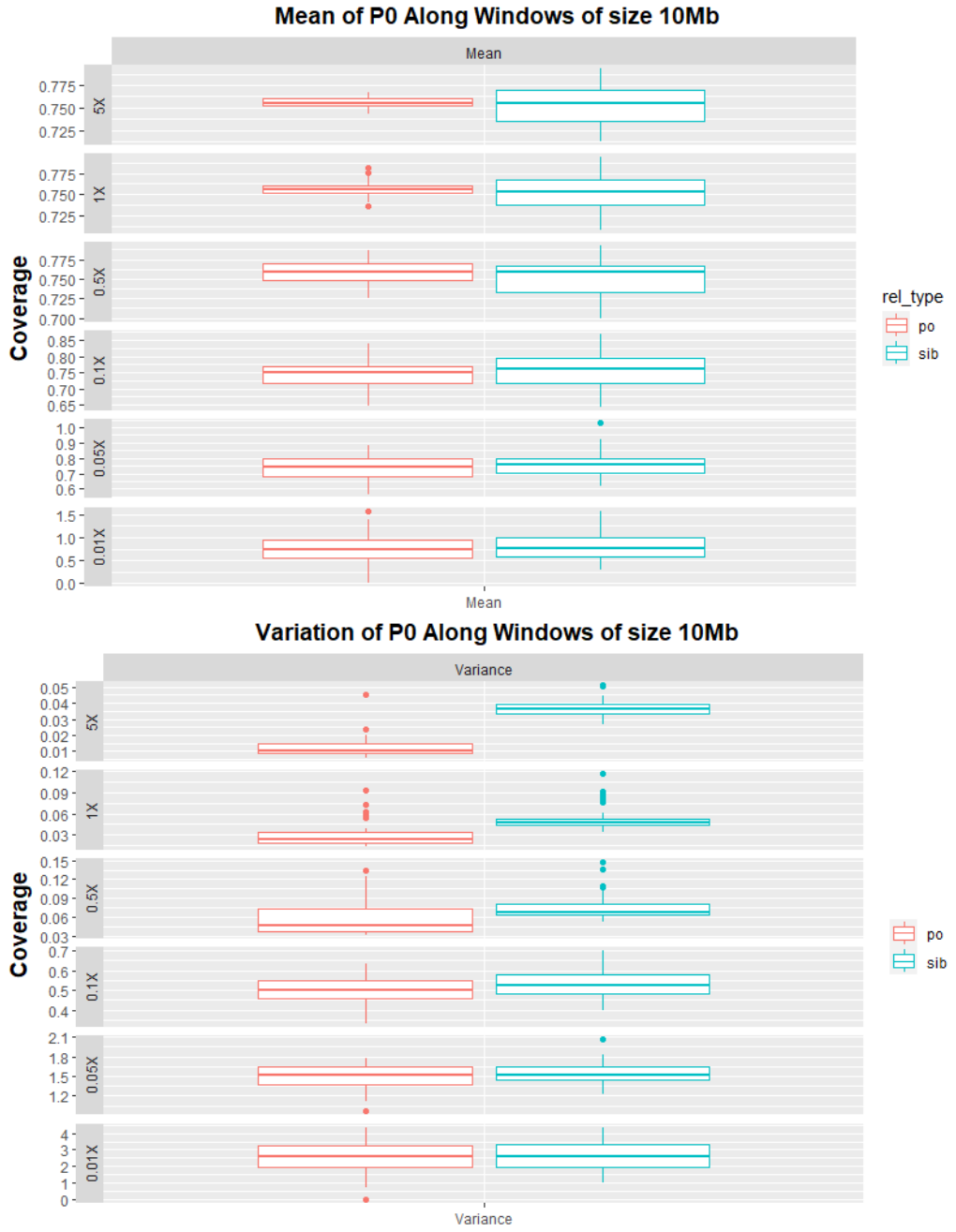
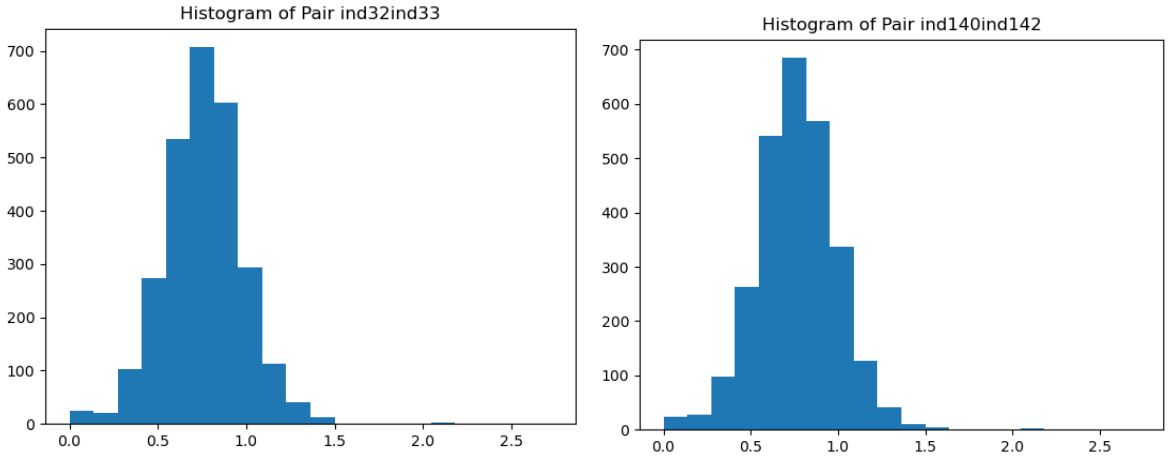


Figure S1: The mean and variance of normalized P0 values of windows of size 1Mb and 10Mb. Similar to Figure 5, there is a clear separation of variance values for high coverages, and variation seem to be increasing and converging to the maximum value of 4, as the coverage drops, preventing to determine a clear cutoff value that could be used to classify first-degree relationships. As expected, and similar to Figure 5, the mean values are the same for these two different categories in different window sizes.

Appendix D – Histogram Plots from Extra Window Sizes

1Mb – Parent-offspring pairs:



1Mb – Sibling pairs:

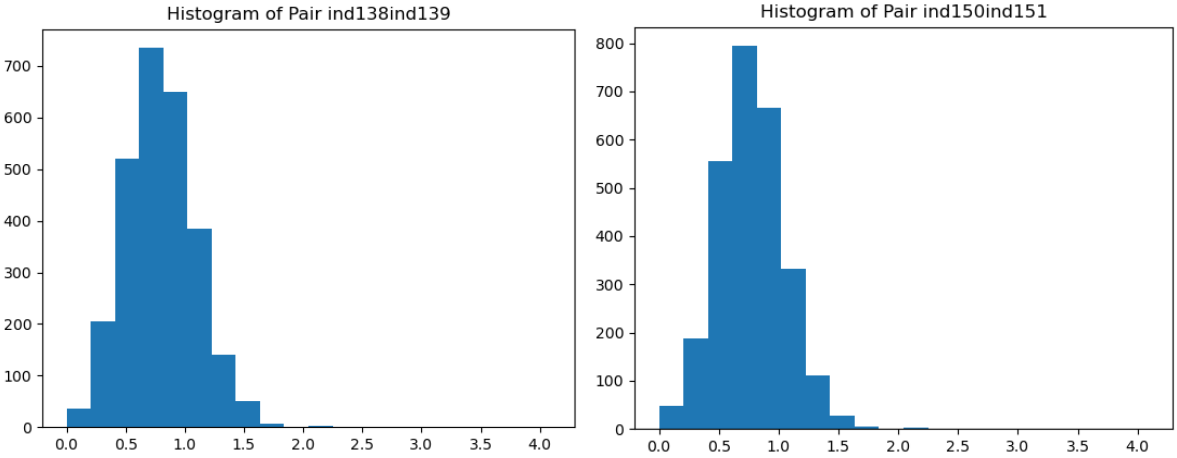
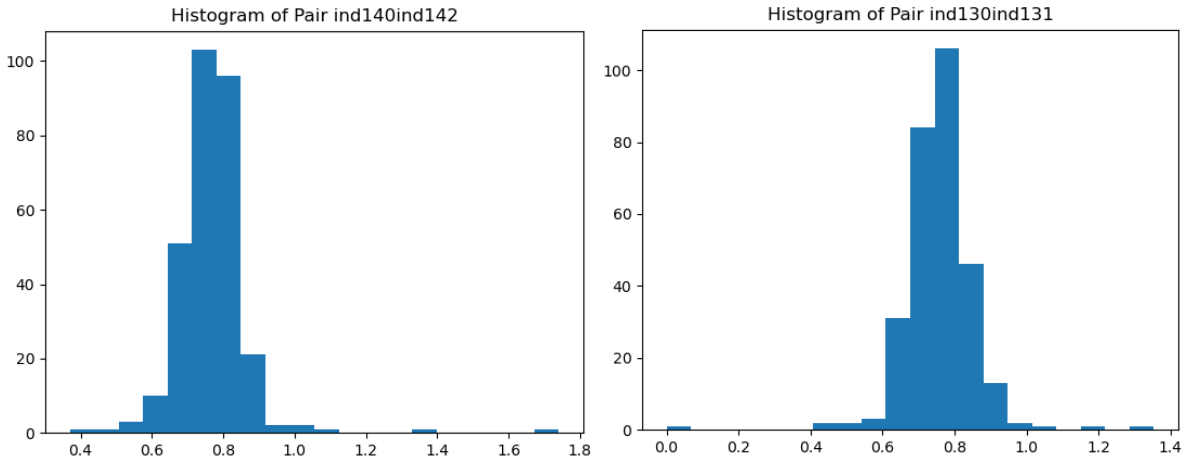


Figure S2: The histogram plots of parent-offspring and sibling pairs of windows of size 1Mb. The parent-offspring histogram starts to look like the ideal plot, the normal distribution, however, the sibling plot is still far away from the histogram plot with three peaks.

10Mb – Parent-offspring pairs:



10Mb – Siblings:

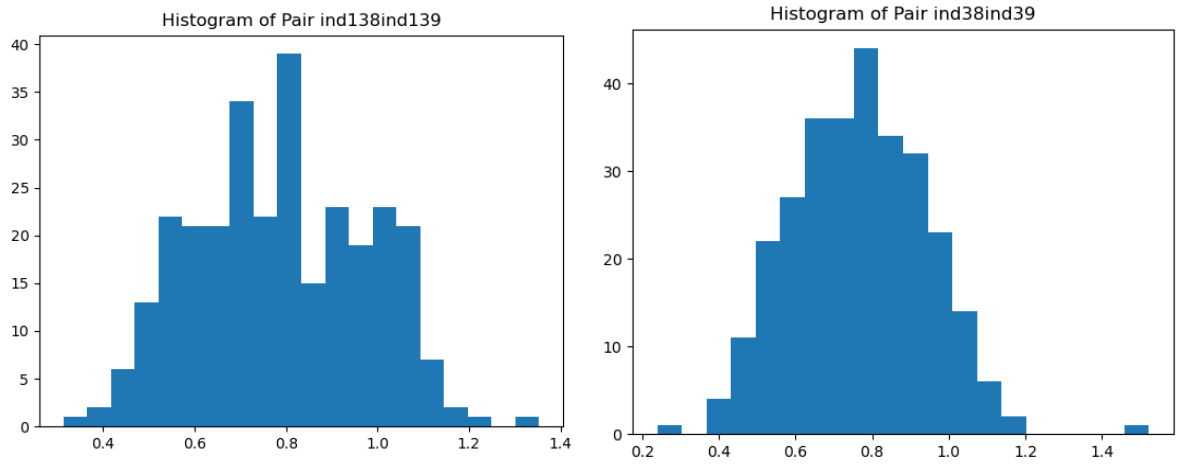


Figure S3: The histogram plots of parent-offspring and sibling pairs of windows of size 10Mb. Parent-offspring plots continue to look like the ideal plot of normal distribution, and sibling plot looks more like the ideal plot with three peaks, however, the areas in identical interval and unrelated interval do not have a clear peak to be identified easily, yet.