Estimating the age of runs of homozygosity in silico

Date: 26th of August 2024
Author: Beatriz Antunes
Supervisors: Anders Albrechtsen and Cindy Santander

Abstract

Runs of homozygosity (ROH) are defined as contiguous stretches of homozygous genotypes within an individual's genome. They provide valuable insights into population history, inbreeding and genetic diversity. The aim of this study is to address the inaccuracies in the existing literature concerning the distribution of ROH age, which is calculated as the number of generations since the ROH was formed. Four distinct demographic scenarios were employed in coalescent-based simulations to generate ancestral trees in which the ROH of sampled individuals were identified, and their size and age were estimated. The results demonstrate a strong negative correlation between the size of ROH segments and their age. Moreover, a comparative analysis using cumulative plots across different demographic scenarios highlights the significant influence of population history on the formation and persistence of ROH.

Keywords: Runs of homozygosity, ROH, Age Estimation, Simulations, Demographic Scenarios, Population Genetics, Inbreeding, Genomic Analysis

1 Introduction

The human genome contains runs of homozygosity (ROH), which are defined as continuous segments of homozygous base pairs within the genome of a diploid individual[1]. These arise when an individual inherits identical segments of DNA from both parents. The analysis of these regions provides powerful insights into the genetic history of a population.

The length of a run of homozygosity is measured in base pairs and serves to delineate the initial and final positions of the continuous stretch of homozygous genotypes. The presence of a longer ROH may be indicative of recent common ancestry, given that there has been less opportunity for recombination. The primary cause of ROH formation is inbreeding[2], defined as the mating of individuals with a close genetic relationship. The probability of ROH formation is elevated within small, isolated or bottleneck populations. At each mating event through the generations, two diploid cells, derived from each parent, undergo a process of transformation into haploid cells. These cells are then fertilised, resulting in the creation of a diploid chromosome[3]. This process is referred to as meiosis and, through a series of iterations, is responsible for the fragmentation of the existing ROH into smaller segments. Consequently, parents with distant relationships will transmit smaller homozygous chromosome segments to their offspring. However, there may be some exceptions due to low recombination rates in specific genomic regions. By analysing the difference between the sum total length of ROH (SROH) and the total number of ROH (NROH), it is possible to gain insight into the demographic origin of the population[2].

Direct empirical testing of evolutionary models requires the undertaking of extensive pedigree studies. The fact that genetic changes are slowly inherited over time, the need for a sufficiently large population and the need for multiple replicates to capture the stochastic nature of evolution result in a demand for an excessive amount of resources[4]. As a solution, computer simulations can be used in evolutionary biology. Forward simulators are more computationally intensive because all individuals are simulated from a given initial state, such as $SLiM$[5]. For this project, a coalescence-

based simulation was chosen as it is more efficient to simulate the coalescence events backwards. This method only traces the lineages of the individuals from the present (time 0).

ROH has been the subject of several investigations in the context of population genetics and disease risk, given its association with recessive genetic disorders. Despite the growing interest in ROH, there is a significant challenge in research regarding the accuracy of the distribution of the age distribution of ROH, as the formula $g = 100/(2rL)$ is being employed [6], where g represents an estimate of the time in generations until a common ancestor is shared by parental lineages, r denotes an estimation of the average recombination rate, and L is the length of ROH. The demographic characteristics of the population are expected to influence the age of the segments, yet this expression does not account for this factor.

Aim

This study aims to address this issue by investigating the correlation between ROH size and age in silico and how the demography of the population influences the distribution of ROH ages. The initial hypothesis is that there will be a strong negative correlation between age and length of runs of homozygosity, as well as a clear influence of the demographic model on the distribution of age.

In order to achieve these goals, the ancestry of genomes for four populations with distinct demographic histories (large, small, bottleneck and expanding) will be simulated backwards in time in accordance with the Wright-Fisher model. The simulated chromosomes will be autosomal, and ROH will be identified by comparing the chromosomes of the sampled diploid individuals. The relationship between length and age of these segments will be investigated using density plots. To explore how ROH age distribution varies with size, ROH will be divided into categories according to size intervals, and the frequency of the ages within each interval will be observed. Furthermore, cumulative distribution plots will be generated to compare the impact of different demographic scenarios on ROH age patterns.

2   Methods

Msprime is a coalescence-based method employed in population genetics for the purpose of simulating the ancestral history of the genome of sampled individuals[7]. The process terminates when all samples have coalesced to a single common ancestor. The data generated can be subjected to analysis using *tskit*, the tree sequence toolkit[8]. This library offers a range of functionalities for genetic statistics, including the calculation of averages across sites. These can be determined either over a single sample set (one-way method) or across multiple sets (multi-way method). The software is a Python package which by default assumes a single randomly mating population of fixed size[7]. It would be unrealistic to ignore population structure, size variation, migration, mutation, and selection. Therefore, *msprime* allows the user to define a demographic model. Two constant size diploid populations were modelled, a small one of size 200 (Small) and a large one of size 2000 (Big). To define an expanding population, an initial size of 2000 was set for the present with a growth rate of 0.03. With regard to the bottleneck population, an instantaneous bottleneck was added 25 generations ago with a strength of 200, indicating that 25 generations ago there was a theoretical collapse of the previous 200 generations[7].

To simulate each population, 200 diploid individuals were sampled, which corresponds to 400 autosomal chromosomes with a length of $1e^8$ base pairs and a recombination rate of $1e^{-8}$. The final option employed for this function is the model under which the ancestral tree was generated. The Discrete Time Wright-Fisher (DTWF) model was selected for the two constant size populations and the expanding one. The Wright-Fisher model simulates generation by generation backwards in time, thereby capturing the complexities of recombination and genetic drift[9]. The Wright Fisher model is based on certain assumptions, including no overlapping generations. The simulations were also modelled without mutations, migration and selection. However, the instantaneous bottleneck population is not supported by this model, and therefore the Hudson model was used instead. This classical coalescent with recombination model is an efficient continuous
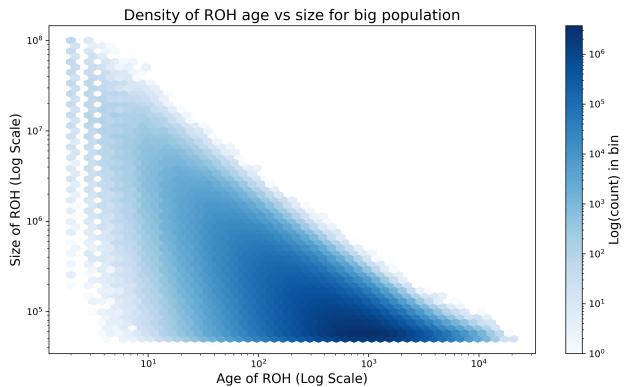
time model[7], and the usage of a different model does not affect the conclusions.

To identify the ROH segments, a comparison was conducted between the chromosomes pairs. In the absence of an established definition of the requisite length of a homozygous segment for classification as a ROH, the minimum span was set at 5000 base pairs. To ensure that all positions within the segment were of the same age and to substantiate that the age of the segment could be extrapolated by calculating the age of a single position, a supplementary analysis was conducted. By analysing the tree at a specific genomic position, the time in generations to the most recent common ancestor for that pair of chromosomes was determined. The length of the homozygous segments was then deduced based on the knowledge of their right and left genomic positions. This process was carried out on 10000 different seeds to ensure the generation of diverse simulations of phylogenies and to allow the reproduction of the results if necessary.
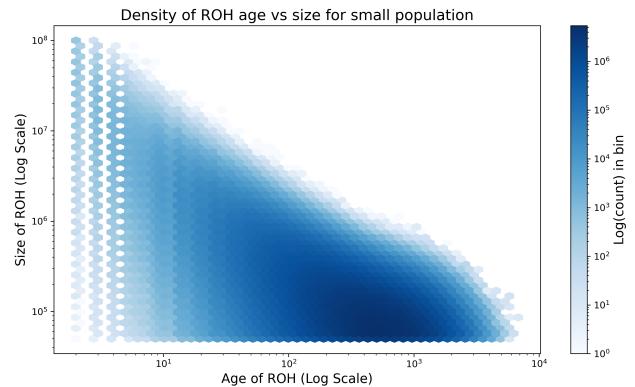
3   Results
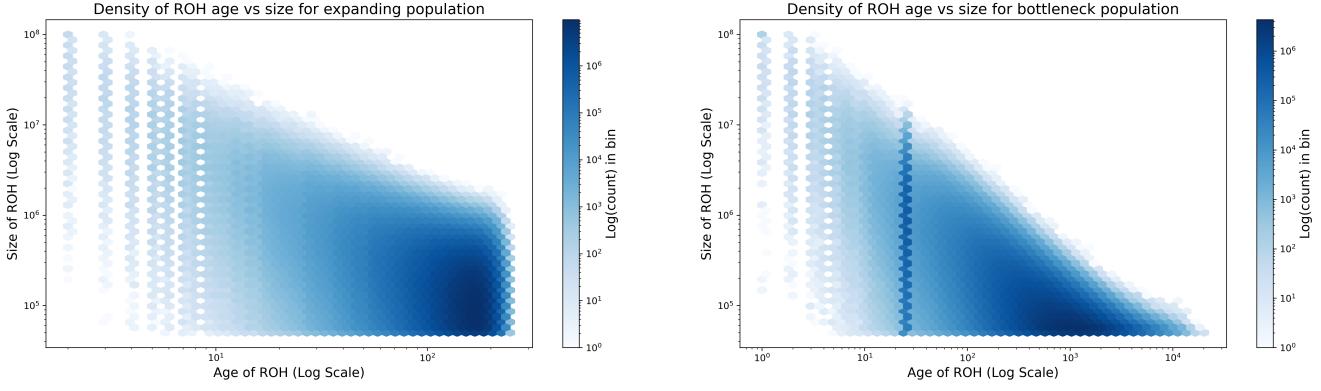
In order to analyse the correlation between age and length of runs of homozygosity for all simulated diploid individuals, a series of visual and statistical analyses were conducted. Age was plotted against ROH size in four separate density plots, one for each demography (Figure 1). The use of density allows for the highlighting of areas with higher concentrations of data points, thereby facilitating the identification of the coordinates at which the relationship between age and ROH size was most pronounced for each demographic group. In addition, the logarithmic scale was used to capture the relationship between the variables most clearly. On a normal scale, the data would manifest an exponential shape, making it challenging to identify trends. The utilisation of a logarithmic scale enables the linearisation of the relationship between the variables, facilitating the interpretation. Given the extensive range of magnitudes encompassed by the variables, the log scale compresses the data, enhancing its visualisation and comparison.

(a) Density plot between ROH age and size of 200,000 diploid individuals from a big population of size 2000.

(b) Density plot between ROH age and size of 200,000 diploid individuals from a small population of size 200.

(c) Density plot between ROH age and size of 200,000 diploid individuals from a expanding population with a initial size of 2000 and a growth rate of 0.03.

(d) Density plot between ROH age and size of 200,000 diploid individuals from a bottleneck population where the bottleneck event happened 25 generations ago with a strength of 200.

Figure 1: ROH age is in generations and ROH size is in base pairs. Both axes are plotted on a logarithmic scale that has been adapted using the formula $\log(x + 1)$ to handle values of zero.
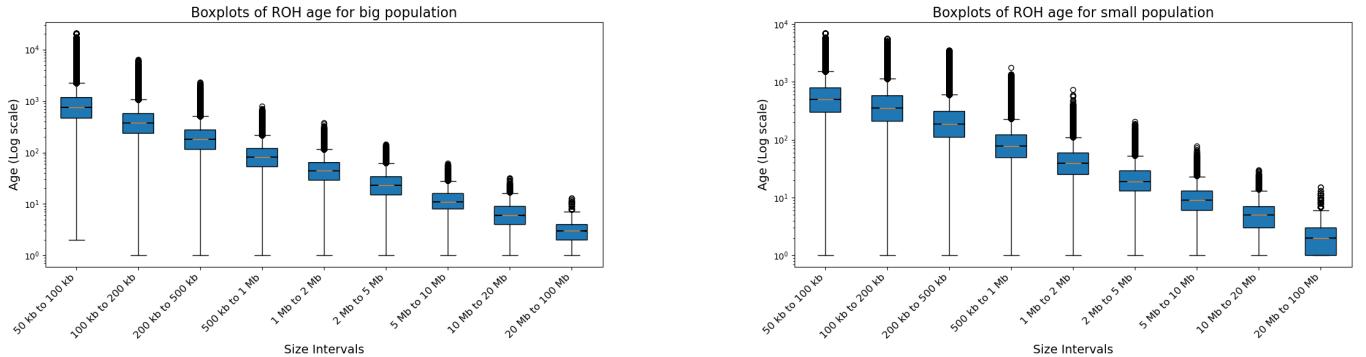
The total length and number of ROH were calculated for the 200,000 diploid individuals simulated, segregated by demographic group (Table 1). NROH and SROH are influenced by factors such as population size, mating system and demographic history. Therefore, calculating these values per demography provides insights into inbreeding patterns and genetic structure in populations.

Demography	NROH	SROH
Big	$1.63e^8$	$1.69e^{13}$
Small	$5.53e^8$	$8.41e^{13}$
Expanding	$5.69e^8$	$9.14e^{13}$
Bottleneck	$1.78e^8$	$2.37e^{13}$
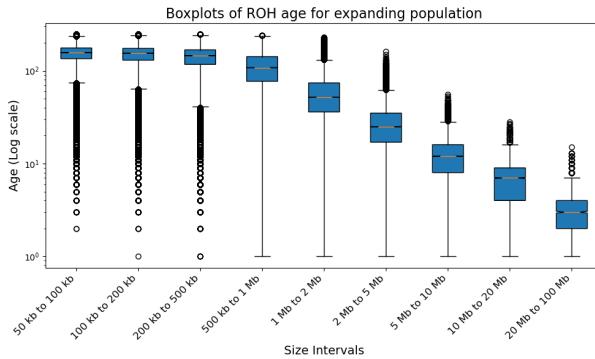
Table 1: NROH (Number of ROH) and SROH (Sum of ROH) measure in base pairs.

All homozygous segments were clustered according to their size, and histograms were generated to analyse the age patterns of the segments according to their length. As a better way of presenting the key statistics for these intervals according to demography, a table with mean, mode, median, range and 90% range was created. This provides a clear numerical comparison of the age distribution across ROH size intervals. The histograms and tables can be found in the Appendix section.

In order to facilitate the comparison of the age distributions across different ROH size intervals, box plots were employed as a visual representation of the dispersion or agglomeration of the data, thereby providing evidence of the differences between the ranges of values (Figure 2). The box plot offers a concise overview of the data, encompassing the median (50%), quartiles (25% and 75%), and potential outliers. The median represents a mesure of central tendency, whereas the interquartile range, as indicated by the size of the box, is a measure of dispersion. Outliers are visually represented as individual data points. A comparison between box plots allows insights to be gain into the distribution of age across different ROH sizes. Once more, the logarithmic scale has been employed on the y axis to more effectively illustrate the pattern and distribution due to the considerable disparity in the orders of magnitude of the values. The whiskers extending from the box in the box plot represent an extension of $1.5 \times IQR$, where $IQR = Q_3 - Q_1$, from the extremities of the box. The whisker from Q_1 is longer than the other one due to the logarithmic scale. As previously mentioned, the same interval between lower numbers will be larger than the same interval for higher numbers.
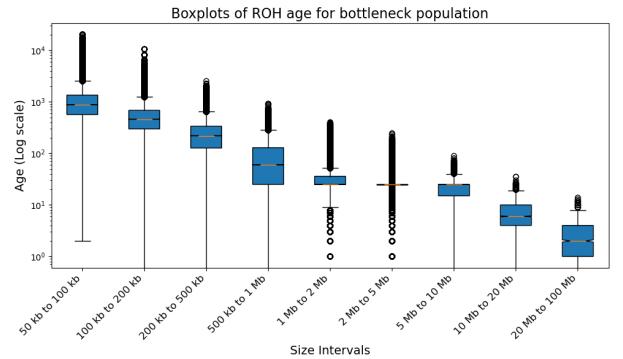
(a) Boxplot of ROH age of 200,000 diploid individuals from a big population of size 2000, categorized by size intervals.

(b) Boxplot of ROH age of 200,000 diploid individuals from a small population of size 200, categorized by size intervals.

(c) Boxplot of ROH age of 200,000 diploid individuals from a expanding population with a initial size of 2000 and a growth rate of 0.03, categorized by size intervals.

(d) Boxplot of ROH age of 200,000 diploid individuals from a bottleneck population where the bottleneck event happened 25 generations ago with a strength of 200, categorized by size intervals.

Figure 2: Boxplots with the ROH segments of 200,000 diploid individuals per demography. ROH Age is in generations and ROH size is in base pairs, where kb is kilobase pair and Mb is megabase pair.

In order to illustrate the impact of demography on the age distribution, cumulative plots have been generated. The y axis represents the cumulative frequency, which is defined as the fraction of ROH with an age that is less than or equal to the number of generations indicated on the x axis. The steeper the slope, the higher the concentration of segments in that age range. This facilitates the identification of the age ranges where demography has a noticeable effect. These plots enable the comparison of the effect of demography on age within each ROH size category and allow for the quantification of the differences in the distributions. The cumulative plots for the minimum, maximum and an intermediate ROH size category are included in the report. For a more comprehensive analysis, the remaining are incorporated in the Appendix section.
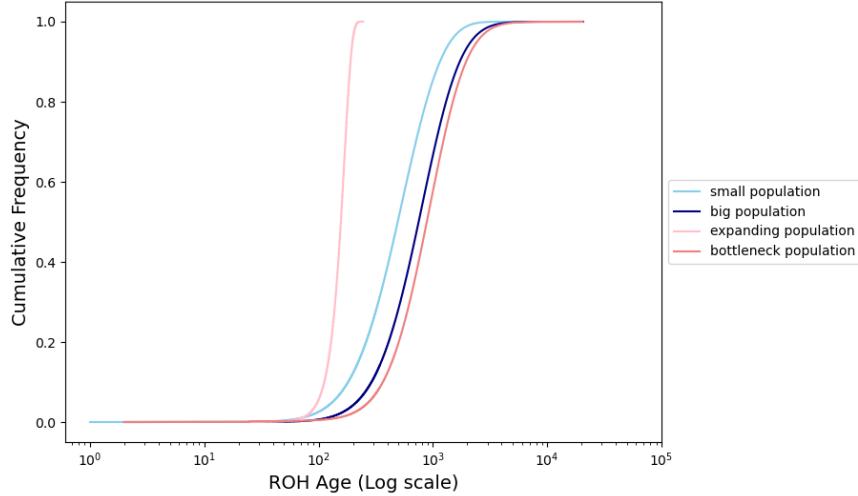
Figure 3: Cumulative plot of ROH age from segments between 50kb and 100kb comparing the four demography.
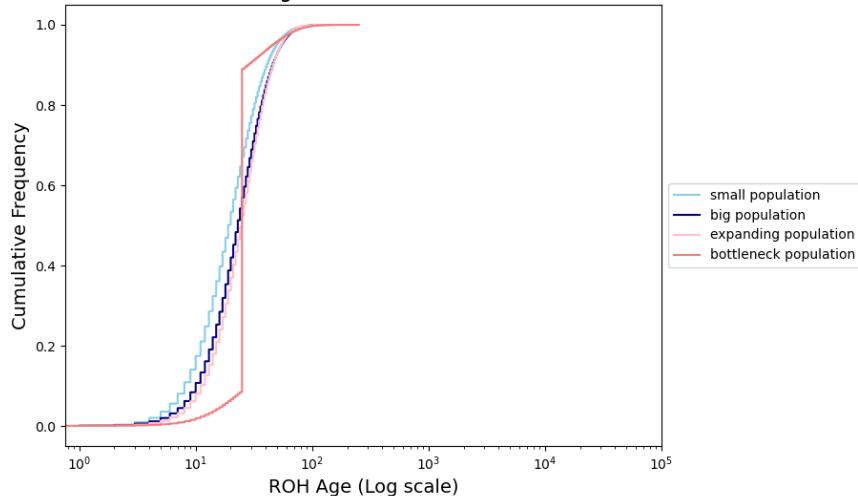
Figure 4: Cumulative plot of ROH age from segments between 2Mb and 5Mb comparing the four demography.
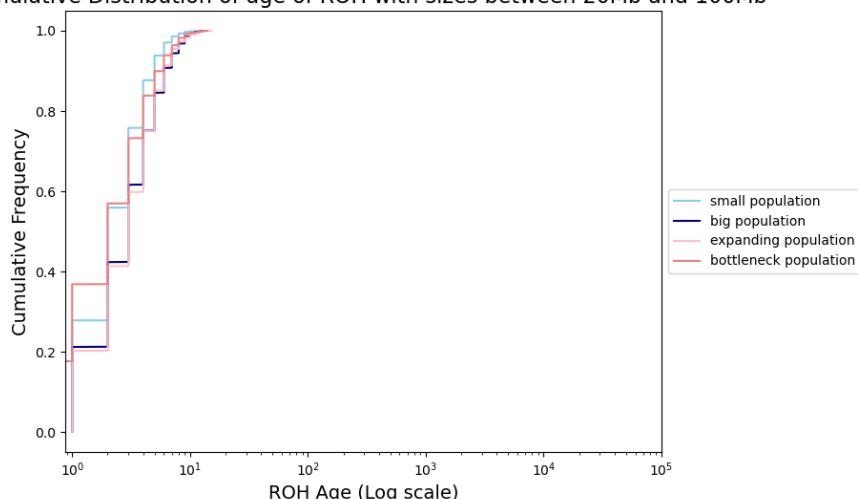

Figure 5: Cumulative plot of ROH age from segments between 20Mb and 100Mb comparing the four demography.

4 Discussion

A visual inspection of the density plots (Figure 1) reveals a clear pattern for each demography. Given that the variables are plotted on a logarithmic scale, it is evident that the rate of change of ROH size with respect to age is not constant. The data points in each plot exhibit a negative slope, a decrease in ROH size as age increases. In this type of plot, regions of a darker hue indicate a higher density of data points. The darker areas in the right-hand side of the four plots are likely to be a consequence of the logarithmic scale employed. The distance on the axis between 10 and 100 is equivalent to 100 and 1000. As a result, even if the current data point density is relatively uniform, the logarithmic scale will appear to exhibit a higher concentration of points in the higher value regions. A comparison of the ranges of both axes reveals that, while ROH lengths vary within the same interval for the four demography, the age distribution has a different range depending on the population. Figure 1c illustrates that the maximum age is much lower in the expanding population compared to the others. This is mainly due to the ancestral history of this population. This population has 2000 diploid individuals in the present and a growth rate of 0.03, so this population has approximately 253 generations. Furthermore, in an expanding population the probability of recombination occurring is increased, resulting in the fragmentation of older ROH segments that are already small. Older ROH are linked to past inbreeding instances. As the population size expands, the likelihood of observing segments of homozygous genotype declines. There is a remarkable difference in the bottleneck population (Figure 1d), which is the presence of ROH with an age of 0 generations. The bottleneck event was recent, 25 generations ago, so the probability for recombination to have happened is low. Given the substantial reduction in the population size, the genetic variation decreased, whereas the probability of inbreeding events increased significantly, as did the number of ROH. In this scenario, new ROH are formed in each new generation, and thus, in the present there are ROH that have just been formed.

With regard to Table 1, the big population exhibits the lowest NROH and SROH, as the probability of genetic recombination is higher, resulting in the breakdown of existing ROH and the formation of shorter and fewer ROH. The second lowest values are observed in the bottleneck population. A bottleneck implies an abrupt reduction in population size, resulting in a significant loss of genetic diversity, a phenomenon known as the founder effect. A reduction in the number of individuals results in an increased probability of closely related individuals mating. This induces an increase in the formation of long stretches of homozygous DNA. The value in the table reveals that the total length of ROH compared to the total number of ROH is proportionally higher in comparison to the other demography. The count remains low due to the recent occurrence of the bottleneck event, which has resulted in the formation of new homozygous segments as a consequence of recent inbreeding. In the small population, it is expected that both NROH and SROH would be high, given the increased probability of inbreeding. In the expanding population, the average number of offspring per individual rises with each generation, leading to a greater prevalence of closely related individuals. However, the increased recombination rate results in shorter segments, but it balances out with new segments that are continuously being formed. So both count of ROH and total length are high.

Upon examination of the box plots (Figure 2), it becomes evident that there is a negative correlation between age and size. As the size of the segments decreases, the age increases. This type of visualisation does not indicate the number of data points within the specified interval. Instead, it depicts the distribution of data. A similar pattern is evident between the big and small populations. Despite these two populations having contrasting differences in NROH, the distribution of ROH age across the various intervals is proportional. The interquartile range of the small population is slightly larger than that of the big population, reflecting the approximately threefold difference in the amount of ROH and the reduced probability of recombination. The top of each box plot contains several potential outliers, while the opposite side contains none. This proves that there is a greater degree of variation at higher ages. A clear distinction is evident in the distribution

depicted in the box plot 2c when comparing the expanding population with the others. The first three size intervals exhibit a similar age range, indicating that these segments are of disparate sizes but were formed contemporaneously. It is crucial to note that the scale of the y axis differs substantially for this population, which has a significant impact on the way these distributions are perceived. The age of the oldest segment is approximately 250 generations old, whereas in the others the maximum age value can be near 20,500. In contrast to the preceding box plots, the potential outliers within the initial three intervals are situated below the first quartile, the lower age values are more dispersed. In the case of the bottleneck population (Figure 2d), there is an evident distinct distribution pattern for the segments between 1Mb and 10Mb in length. These exhibit an extremely narrow interquartile range. The interval from 2Mb to 5Mb is particularly noticeable, as neither the box nor the whiskers are visible and there are several outliers in both directions. The Table 4 allows to infer that 90% of the data points are between 9 to 54, with a median of 25, which corresponds to the generation when the bottleneck event occurred. Therefore, it can be assumed that a significant number of ROH created at that time exist in the present with a size between 2Mb and 5Mb.

The final analysis entailed the creation of cumulative plots that illustrate the influence of the demography on age distribution, thereby addressing the second research question. This plot offers valuable insights that were previously unidentifiable. With regard to the interval from 50kb to 100kb, the expanding population is distinguished by the most pronounced pattern in Figure 3. This is the steepest cumulative line, indicating that approximately 100% of the ROH from the expanding population with a length of 50kb to 100kb are between 100 and 200 generations old. The small population exhibits a lower overall age distribution, whereas the big and bottleneck populations follow an identical curve. Looking at Figure 4, for the interval between 2Mb and 5Mb, the bottleneck distribution is distinguished by its unique shape. The sudden jump occurs at generation 25, which coincides with the abrupt event that created the bottleneck. Consequently, 10% of the segments within this size have an age lower than 25,

whereas approximately 80% have an age equal to 25 generations. A remarkable effect observed in this plot is that the distribution for the other three populations is quite similar. When the cumulative plot for the longer segments (Figure 5) is considered, it becomes evident that the distributions follow the same trend for this size interval. There is a clear approximation of the big and expanding populations, which exhibit slightly older segments compared to the other pair of populations. A comparison of the three visualisations reveals that the age distribution is significantly affected when the segments are smaller, while the distinction is less explicit as the size increases.

The results of the previous analysis sustain that demography and ROH lengths must be considered in order to assess the length distribution and calculate the expected generations.

Future

For future research, a variety of additional parameters can be incorporated or modified in the simulations to encompass a broader range of scenarios. The age of the longest ROH can be extracted, one per individual, and then all plotted in a histogram to ascertain the distribution of the ages. The sequence length can be increased to simulate multiple chromosomes instead of just one. Alternatively, the model can be changed instead of using Discrete Time Wright-Fisher model. The addition of pedigree to the analysis will make it more complex, but more conclusions can be drawn from the study. All these adjustments will enrich the knowledge about ROH.

5 Conclusion

To initiate the research, an in silico analysis was conducted to obtain homozygous genotype segments of autosomal chromosomes from 200,000 diploid individuals. By plotting ROH size against age on a logarithmic scale and analysing the box plots, it was demonstrated that there is a negative correlation between these two variables. An examination of the cumulative plots reveals the effect of the demography. Although the degree of change between the demography for the smallest sizes is notable,

8

the effect becomes more subtle as the size increases. In conclusion, there is a discernible distinction in the relationship of the age according to each size interval, and the effect of demography is also size dependent. It is essential to consider both demography and the clustering of segments according to their size.

6 Acknowledgements

References

[1] Mirna Kirin, Ruth McQuillan, Christopher S. Franklin, Harry Campbell, Paul M. McKeigue, and James F. Wilson. Genomic runs of homozygosity record population history and consanguinity. *PLOS ONE*, 5(11):1–7, 11 2010.

[2] Francisco C. Ceballos, Peter K. Joshi, David W. Clark, Michèle Ramsay, and James F. Wilson. Runs of homozygosity: windows into population history and trait architecture. *Nature Reviews Genetics*, 19(4):220–234, 2018.

[3] KJ Hillers, V Jantsch, E Martinez-Perez, and JL Yanowitz. Meiosis. *WormBook*, 2017:1–43, 2017.

[4] Martin Petr, Benjamin C. Haller, Peter L. Ralph, and Fernando Racimo. slendr: a framework for spatio-temporal population genomic simulations on geographic landscapes. *bioRxiv*, 2023.

[5] Benjamin C. Haller and Philipp W. Messer. Slim 4: Multispecies eco-evolutionary modeling. *The American Naturalist*, 201(5):E127–E139, May 2023.

[6] Nedda F. Saremi, Megan A. Supple, Ashley Byrne, James A. Cahill, Luiz Lehmann Coutinho, Love Dalén, Henrique V. Figueiró, Warren E. Johnson, Heather J. Milne, Stephen J. O'Brien, Brendan O'Connell, David P. Onorato, Seth P. D. Riley, Jeff A. Sikich, Daniel R. Stahler, Priscilla Marqui Schmidt Villela, Christopher Vollmers, Robert K. Wayne, Eduardo Eizirik, Russell B. Corbett-Detig, Richard E. Green, Christopher C. Wilmers, and Beth Shapiro. Puma genomes from north and south america provide insights into the genomic consequences of inbreeding. *Nature Communications*, 10(1):4769, 2019.

[7] Jerome Kelleher et al. msprime 1.0: a modern, efficient, and scalable simulator of population genetics. `https://tskit.dev/msprime/docs/stable/intro.html`, 2023. Accessed: 2024-08-14.

[8] Jerome Kelleher et al. tskit 1.0: Efficient tree sequence processing and analysis. `https://tskit.dev/tskit/docs/stable/introduction.html`, 2023. Accessed: 2024-08-14.

[9] Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E Castedo Ellerman, Jared G Galloway, Ariella L Gladstein, Gregor Gorjanc, Bing Guo, Ben Jeffery, Warren W Kretzschumar, Konrad Lohse, Michael Matschiner, Dominic Nelson, Nathaniel S Pope, Consuelo D Quinto-Cortés, Murillo F Rodrigues, Kumar Saunack, Thibaut Sellinger, Kevin Thornton, Hugo van Kemenade, Anthony W Wohns, Yan Wong, Simon Gravel, Andrew D Kern, Jere Koskela, Peter L Ralph, and Jerome Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3):iyab229, 12 2021.

7   Appendix

Click here to get the code to for the simulations.

Histogram of age from ROH sizes from big population between 50kb and 100kb
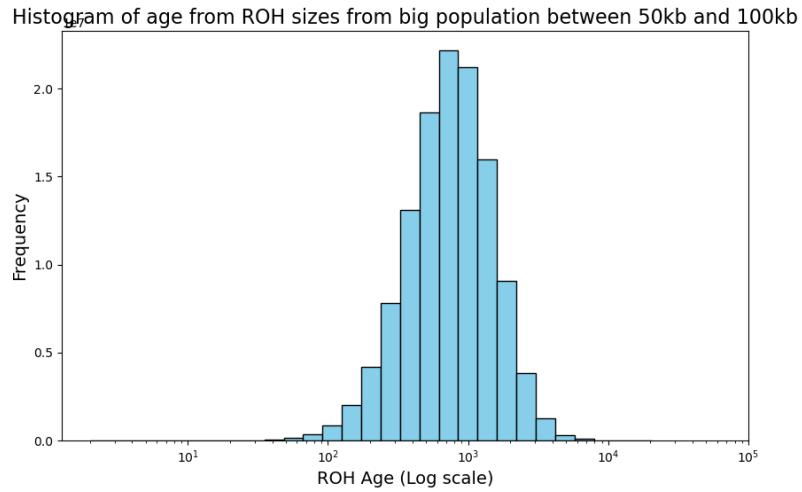
Figure 6: Histogram of ROH age from segments between 50kb and 100kb from big population.

Histogram of age from ROH sizes from big population between 100kb and 200kb
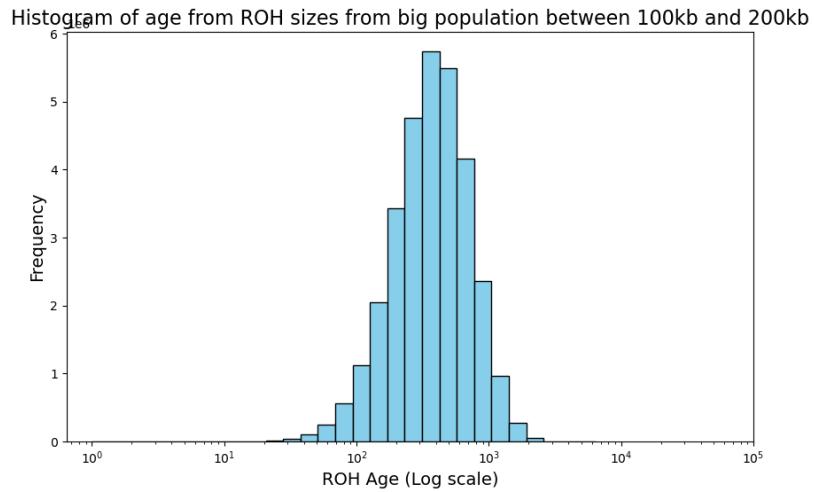
Figure 7: Histogram of ROH age from segments between 100kb and 200kb from big population.
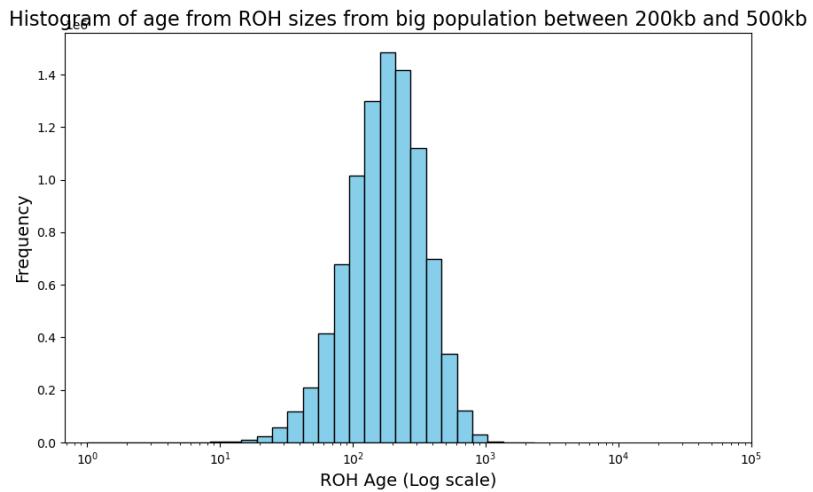
Figure 8: Histogram of ROH age from segments between 200kb and 500kb from big population.
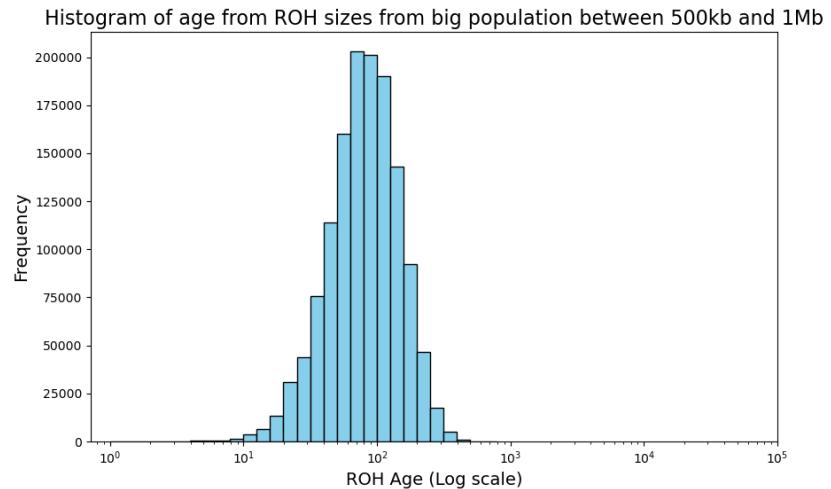
Figure 9: Histogram of ROH age from segments between 500kb and 1Mb from big population.
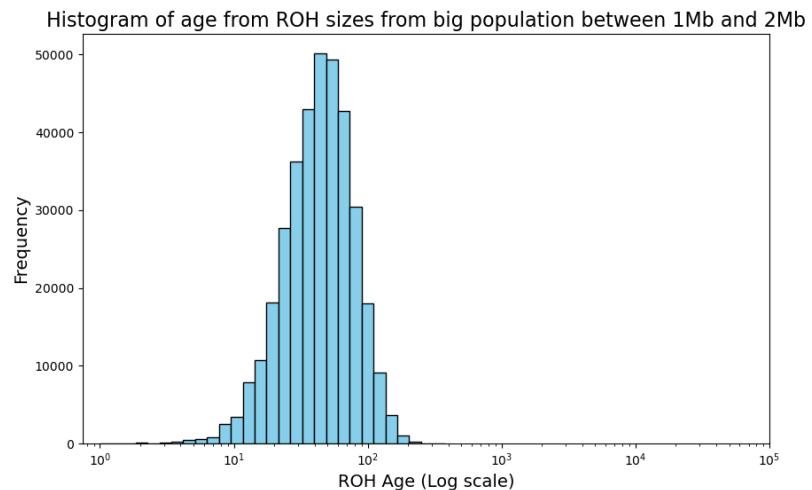
Figure 10: Histogram of ROH age from segments between 1Mb and 2Mb from big population.
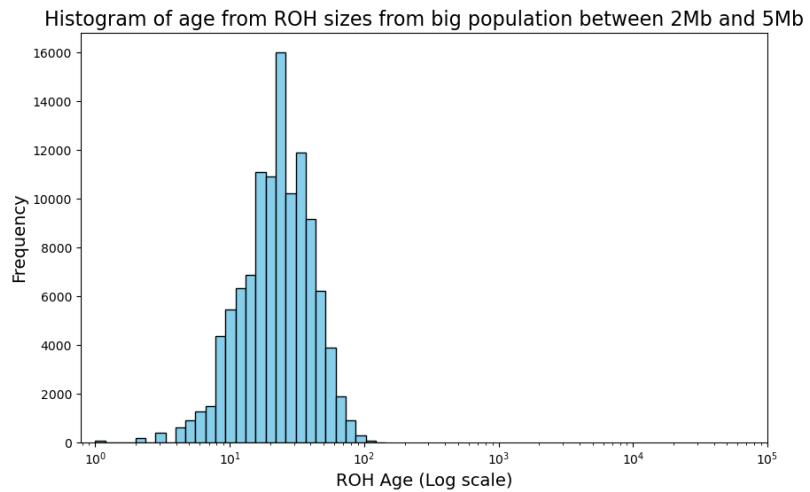
Figure 11: Histogram of ROH age from segments between 2Mb and 5Mb from big population.
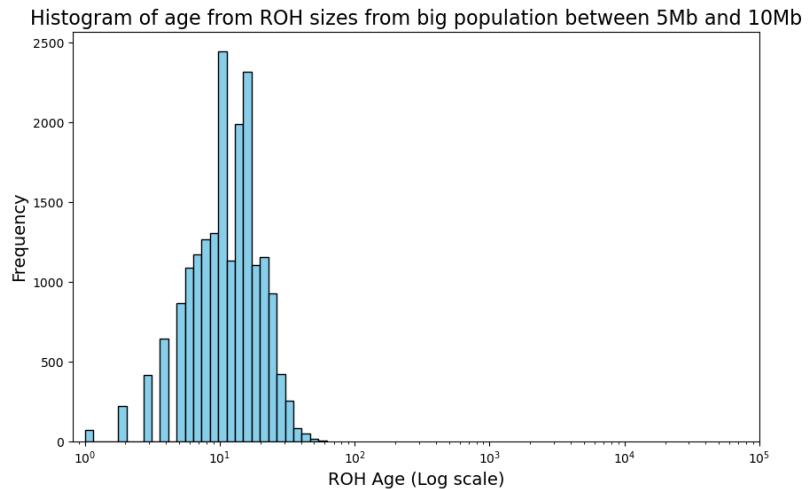
Figure 12: Histogram of ROH age from segments between 5Mb and 10Mb from big population.
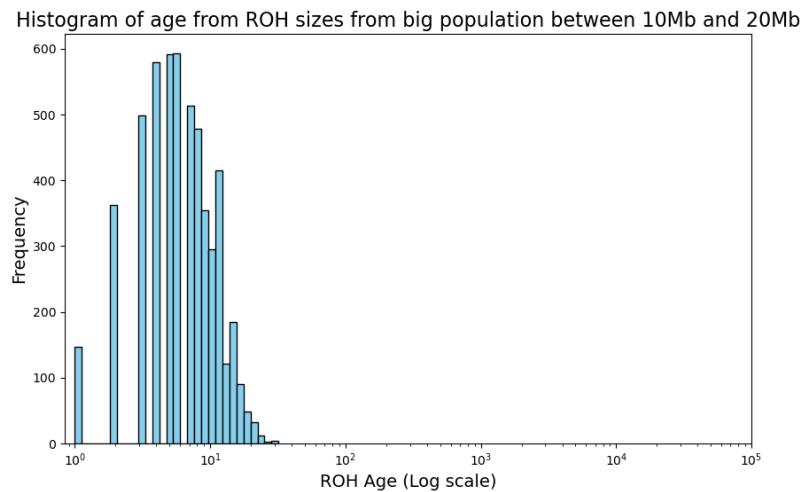
Figure 13: Histogram of ROH age from segments between 10Mb and 20Mb from big population.
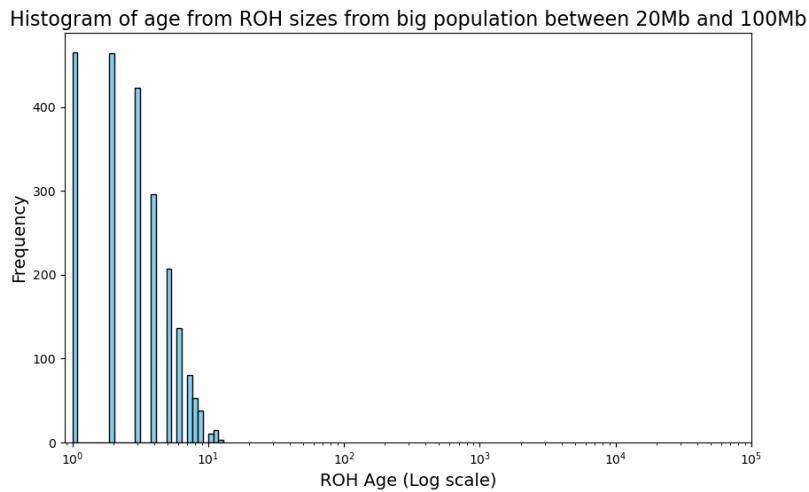
Figure 14: Histogram of ROH age from segments between 20Mb and 100Mb from big population.
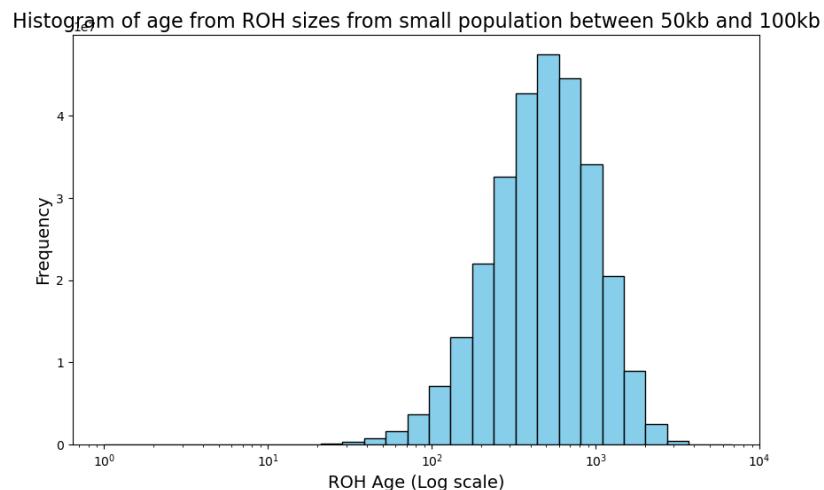
Figure 15: Histogram of ROH age from segments between 50kb and 100kb from small population.
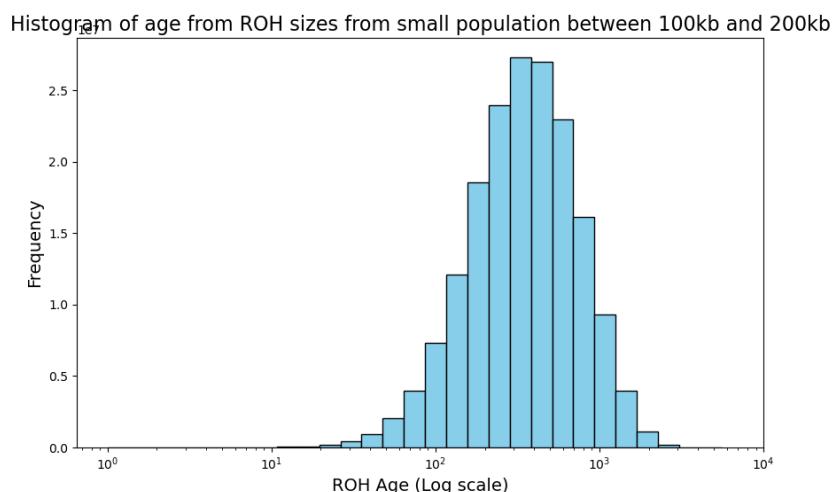
Figure 16: Histogram of ROH age from segments between 100kb and 20kb from small population.
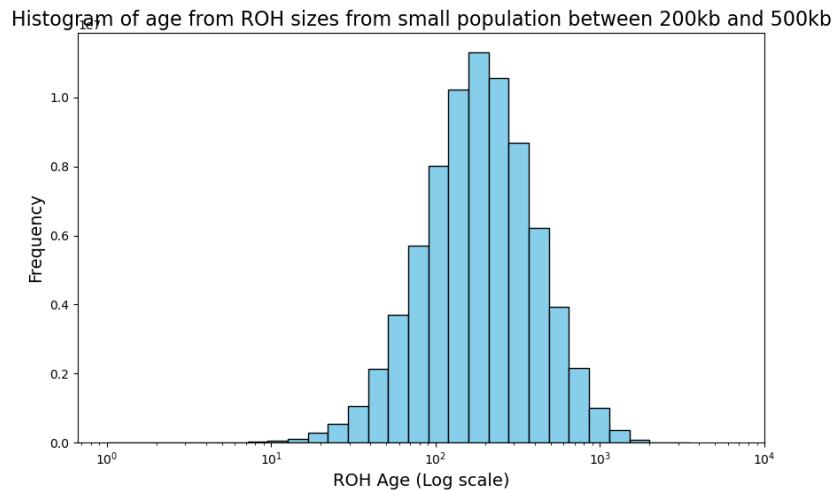
Figure 17: Histogram of ROH age from segments between 200kb and 500kb from small population.
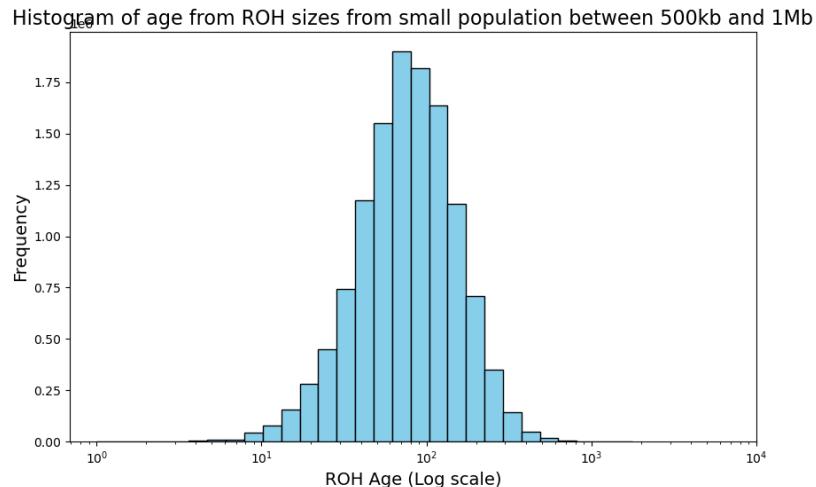
Figure 18: Histogram of ROH age from segments between 500kb and 1Mb from small population.
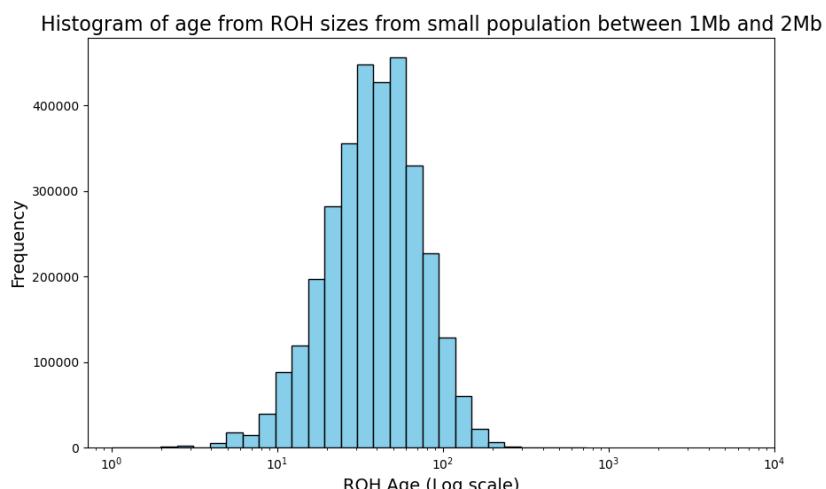
Figure 19: Histogram of ROH age from segments between 1Mb and 2Mb from small population.

14
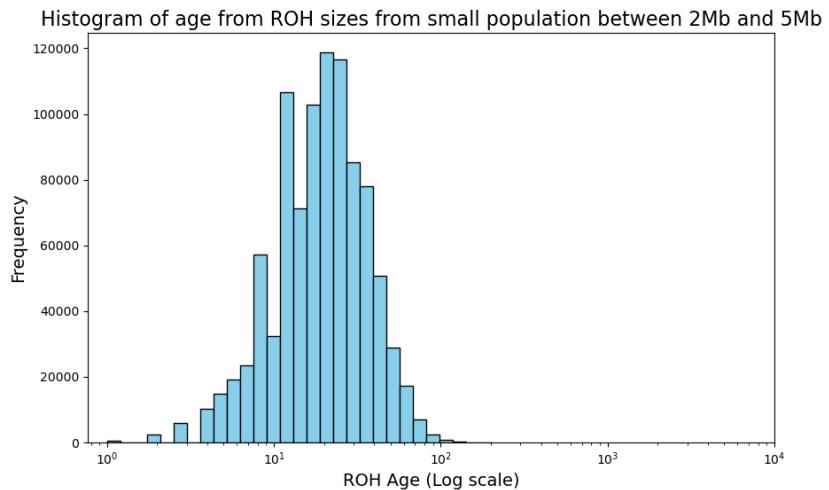
Figure 20: Histogram of ROH age from segments between 2Mb and 5Mb from small population.
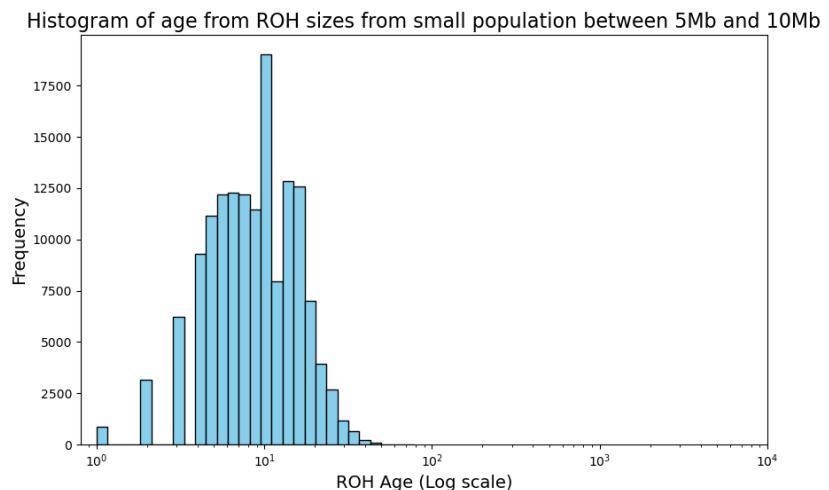
Figure 21: Histogram of ROH age from segments between 5Mb and 10Mb from small population.
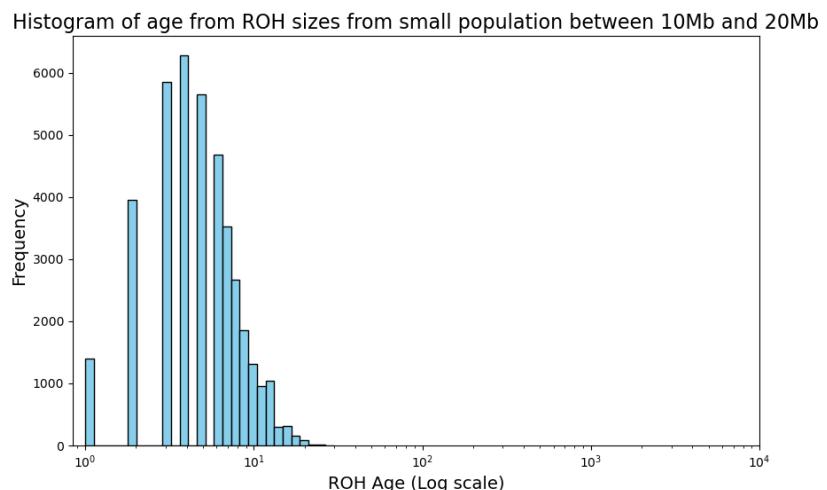
Figure 22: Histogram of ROH age from segments between 10Mb and 20Mb from small population.

Histogram of age from ROH sizes from small population between 20Mb and 100Mb
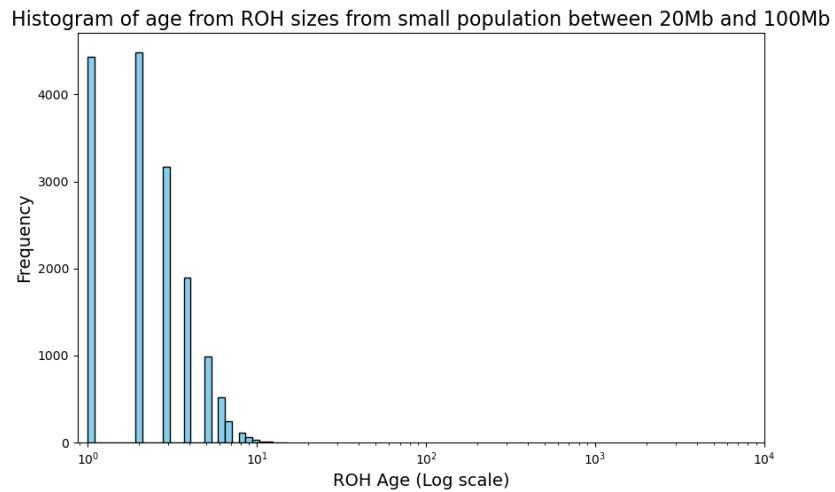
Figure 23: Histogram of ROH age from segments between 20Mb and 100Mb from small population.

Histogram of age from ROH sizes from expanding population between 50kb and 100kb
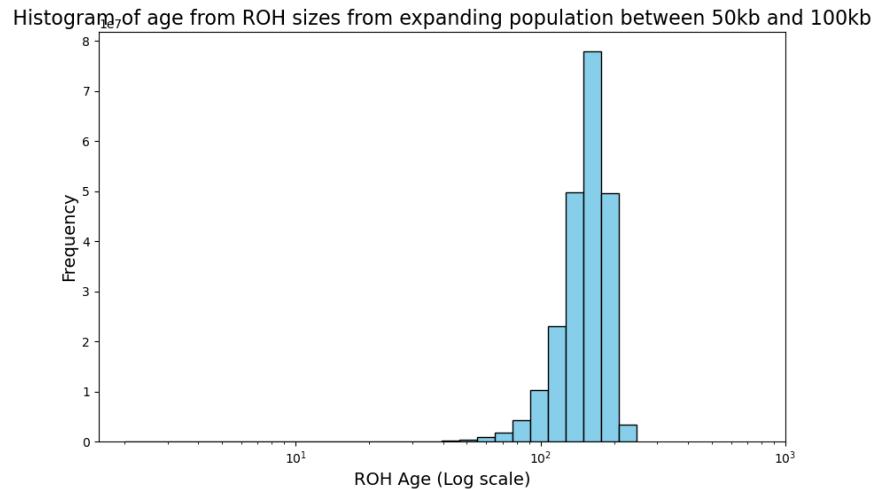
Figure 24: Histogram of ROH age from segments between 50kb and 100kb from expanding population.

Histogram of age from ROH sizes from expanding population between 100kb and 200kb
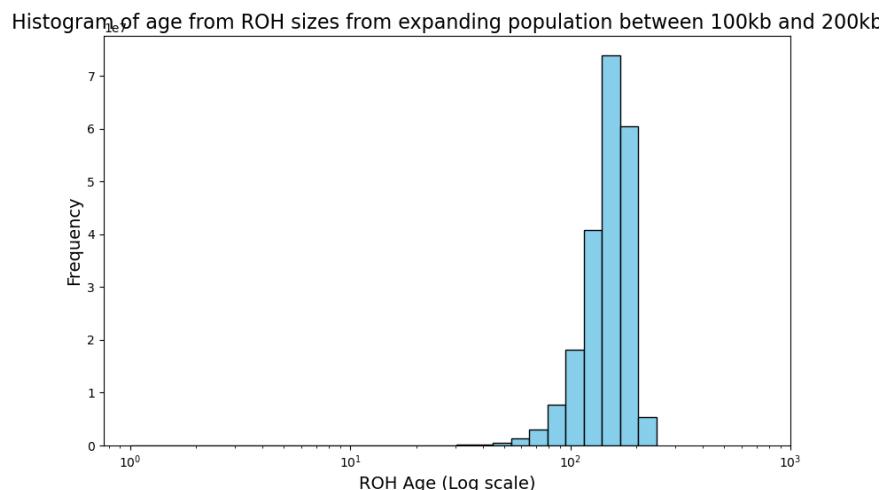
Figure 25: Histogram of ROH age from segments between 100kb and 200Mb from expanding population.
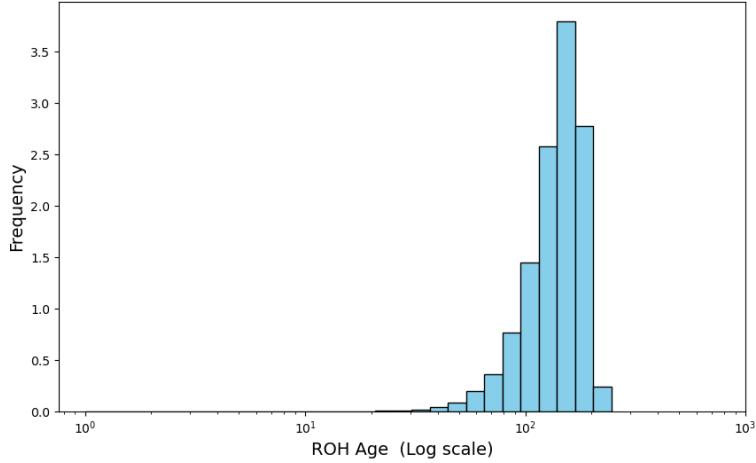
Figure 26: Histogram of ROH age from segments between 200kb and 500kb from expanding population.
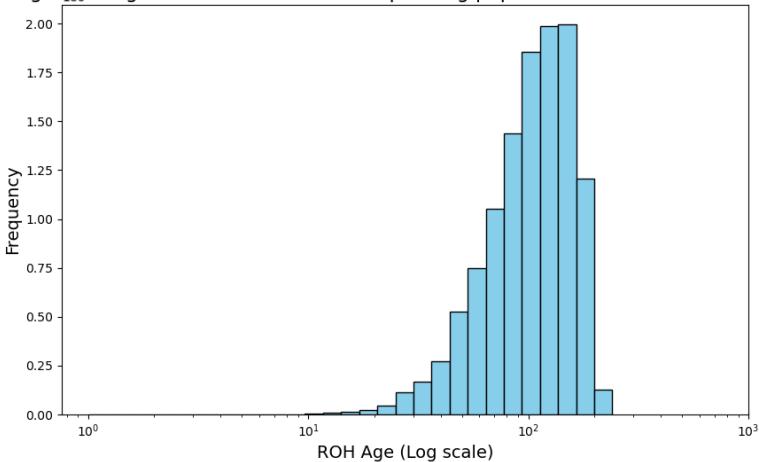
Figure 27: Histogram of ROH age from segments between 500kb and 1Mb from expanding population.
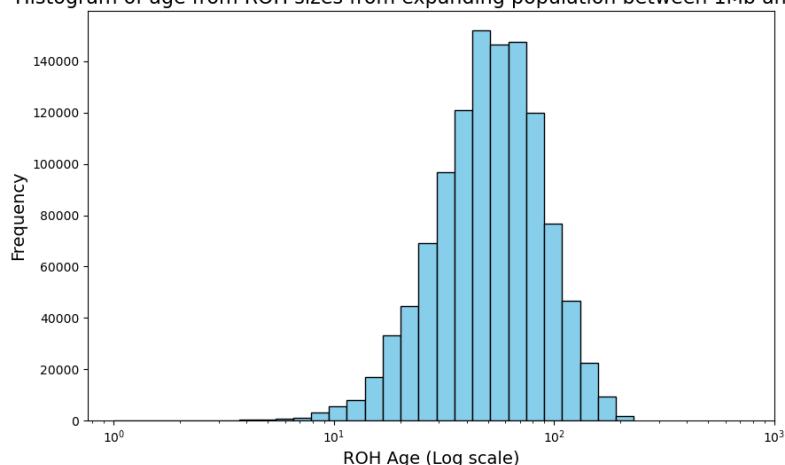
Figure 28: Histogram of ROH age from segments between 1Mb and 2Mb from expanding population.
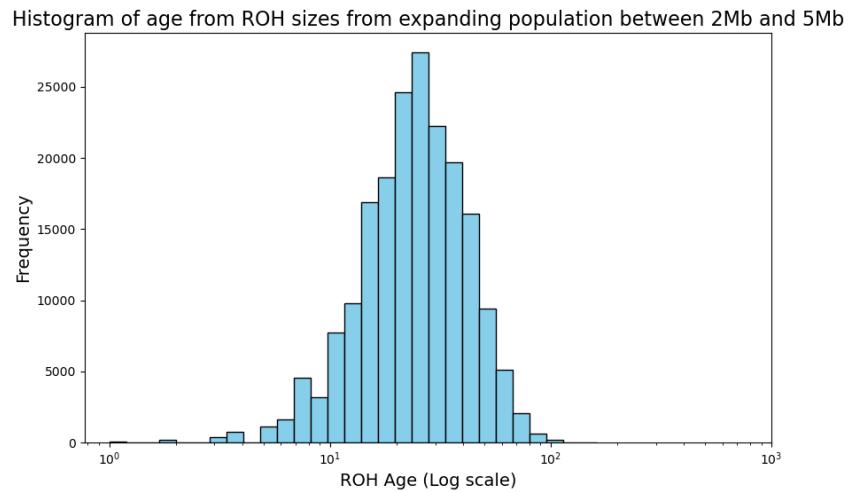
Figure 29: Histogram of ROH age from segments between 2Mb and 5Mb from expanding population.
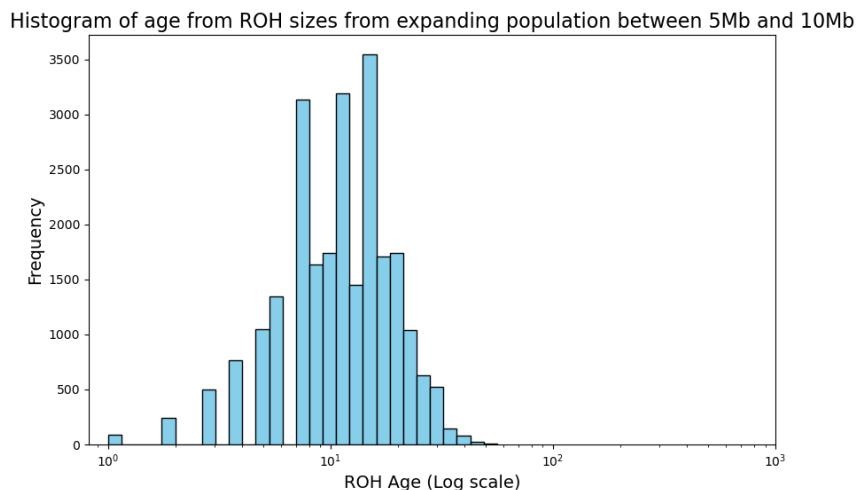
Figure 30: Histogram of ROH age from segments between 5Mb and 10Mb from expanding population.
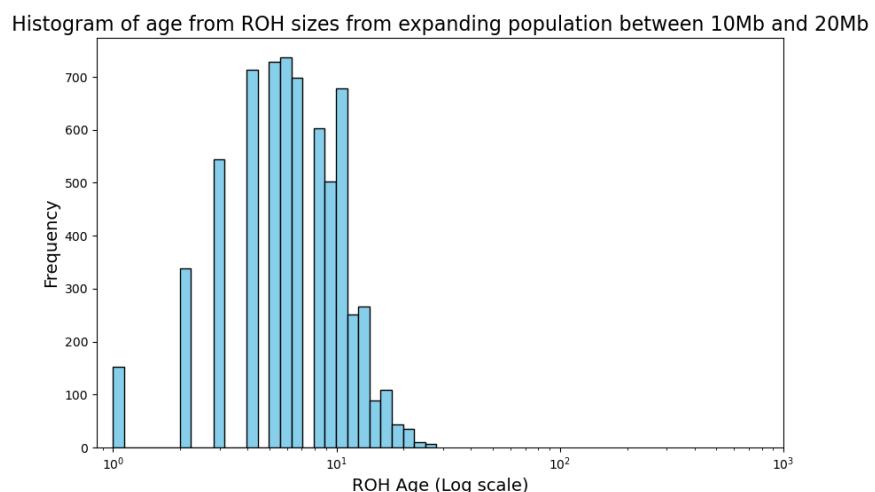
Figure 31: Histogram of ROH age from segments between 10Mb and 20Mb from expanding population.
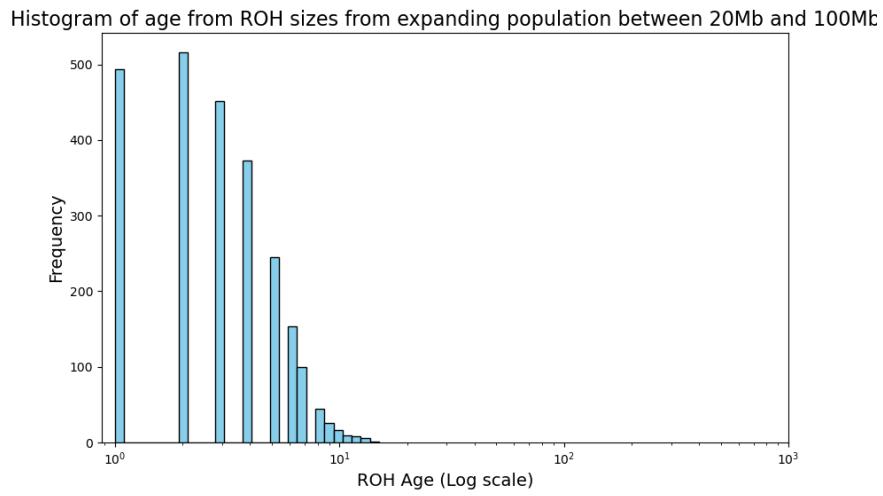
Figure 32: Histogram of ROH age from segments between 20Mb and 100Mb from expanding population.
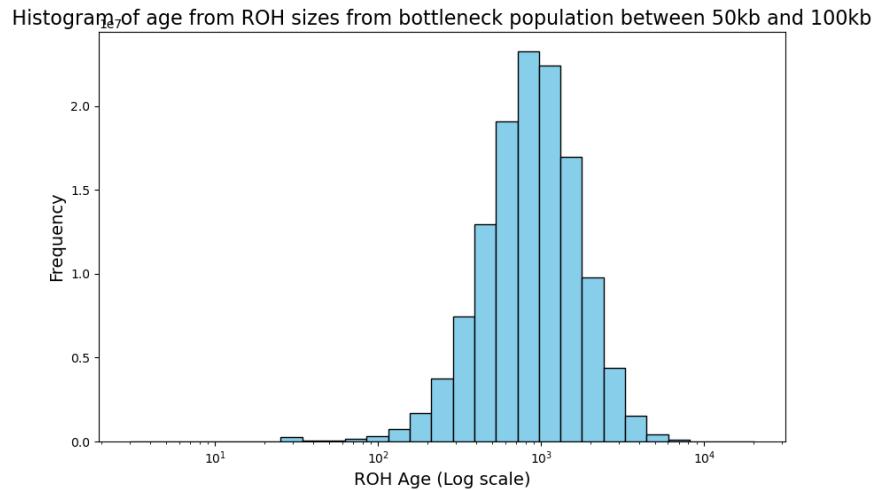
Figure 33: Histogram of ROH age from segments between 50kb and 100kb from bottleneck population.The x-axis is plotted on a logarithmic scale that has been adapted using the formula log(x + 1) to handle values of zero.
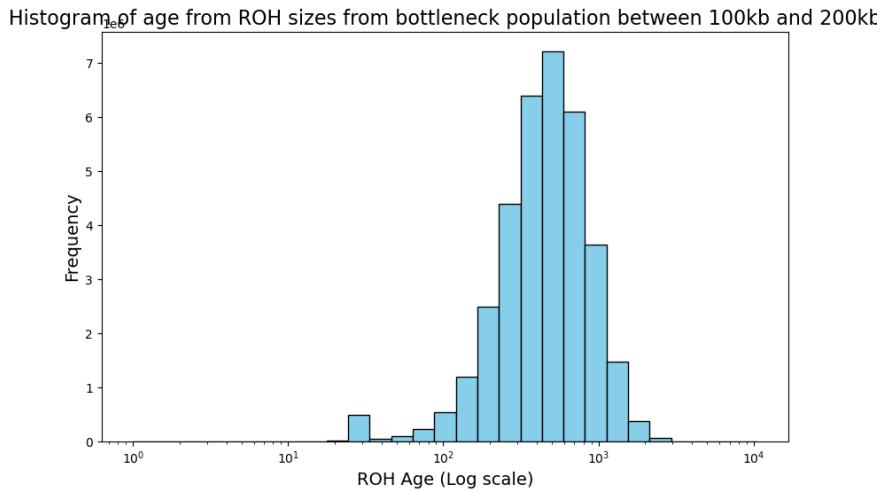

Figure 34: Histogram of ROH age from segments between 100kb and 200kb from bottleneck population.The x-axis is plotted on a logarithmic scale that has been adapted using the formula log(x + 1) to handle values of zero.
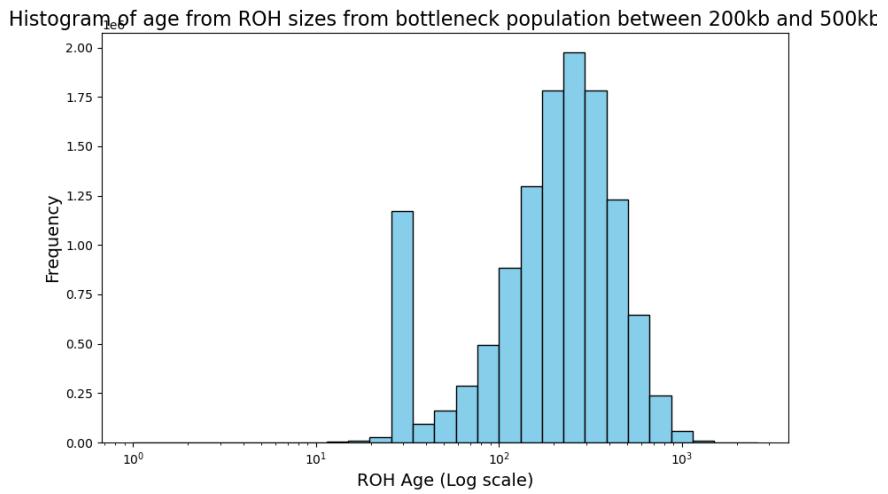
Figure 35: Histogram of ROH age from segments between 200kb and 500kb from bottleneck population.The x-axis is plotted on a logarithmic scale that has been adapted using the formula log(x + 1) to handle values of zero.
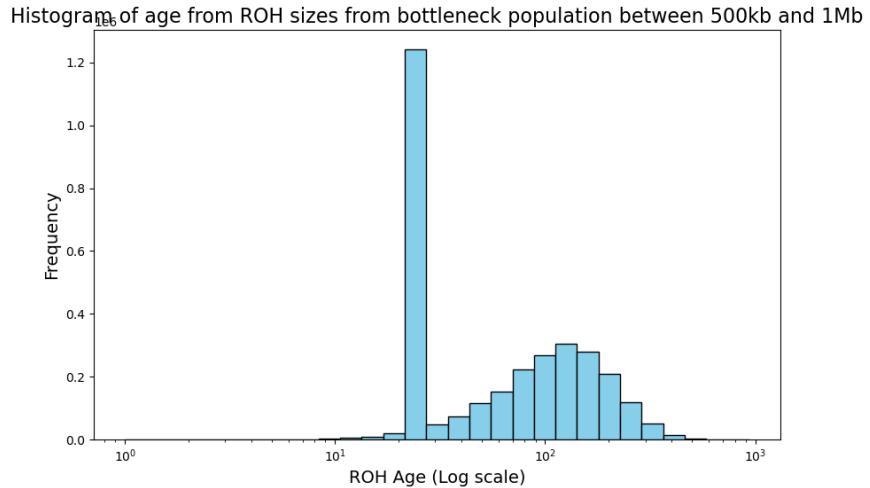
Figure 36: Histogram of ROH age from segments between 500kb and 1Mb from bottleneck population.The x-axis is plotted on a logarithmic scale that has been adapted using the formula log(x + 1) to handle values of zero.
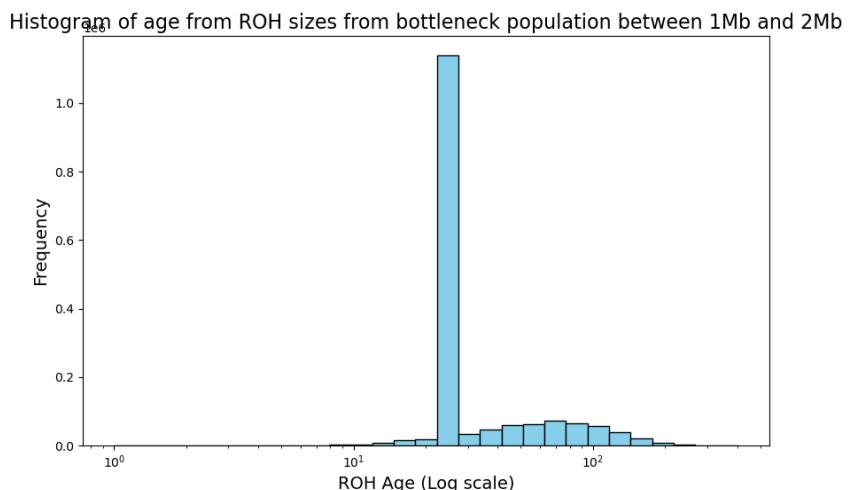

Figure 37: Histogram of ROH age from segments between 1Mb and 2Mb from bottleneck population.The x-axis is plotted on a logarithmic scale that has been adapted using the formula log(x + 1) to handle values of zero.
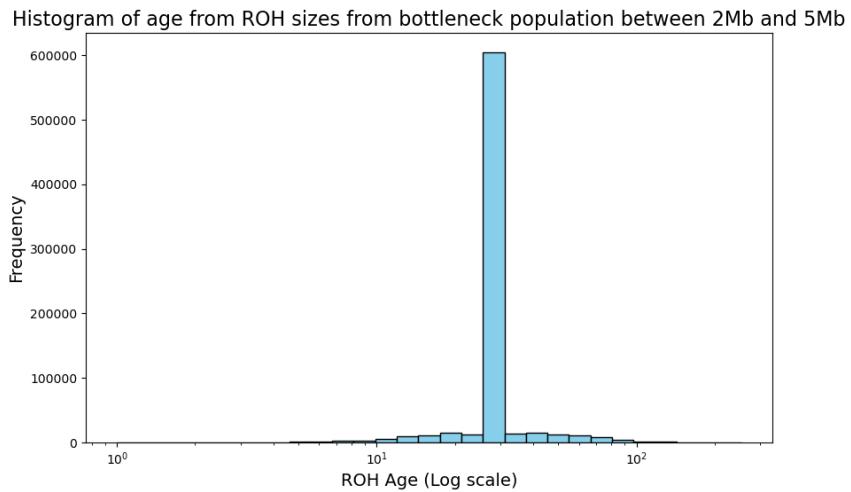
Figure 38: Histogram of ROH age from segments between 2Mb and 5Mb from bottleneck population.The x-axis is plotted on a logarithmic scale that has been adapted using the formula log(x + 1) to handle values of zero.
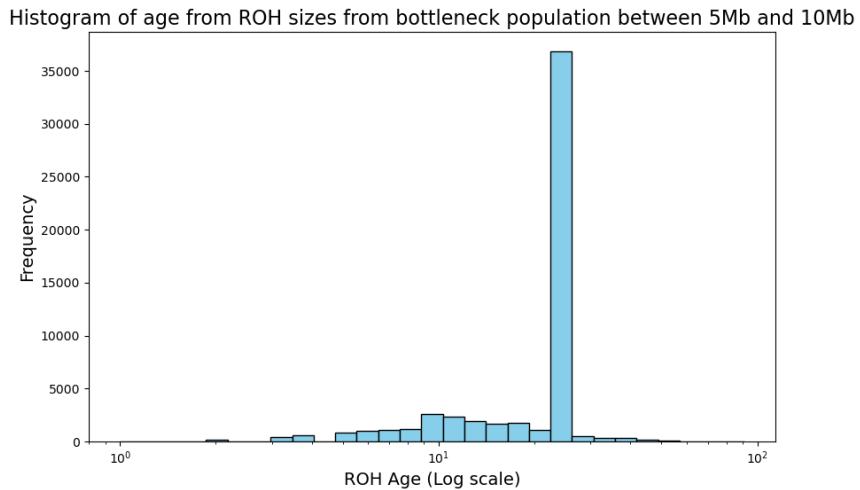

Figure 39: Histogram of ROH age from segments between 5Mb and 10Mb from bottleneck population.The x-axis is plotted on a logarithmic scale that has been adapted using the formula log(x + 1) to handle values of zero.
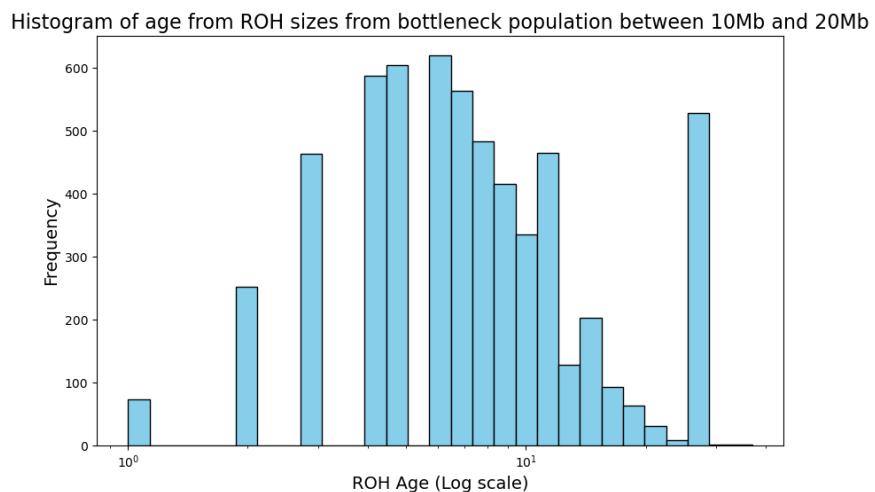

Figure 40: Histogram of ROH age from segments between 10Mb and 20Mb from bottleneck population.The x-axis is plotted on a logarithmic scale that has been adapted using the formula log(x + 1) to handle values of zero.
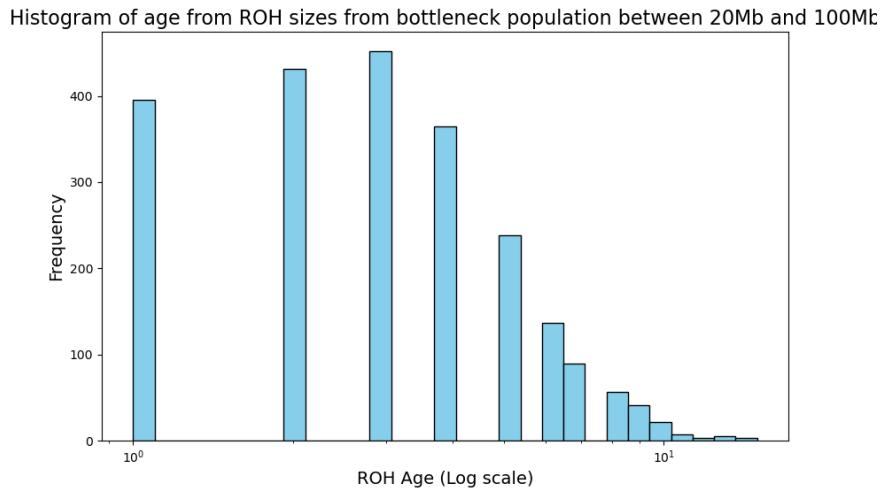
Figure 41: Histogram of ROH age from segments between 20Mb and 100Mb from bottleneck population.The x-axis is plotted on a logarithmic scale that has been adapted using the formula log(x + 1) to handle values of zero.

	Mean	Mode	Median	Range	90% Range
50 kb to 100 kb	913.88	494	751	[2,20550]	[215,2151]
100 kb to 200 kb	441.41	249	374	[1,6430]	[113,997]
200 kb to 500 kb	212.48	117	181	[1,2310]	[56,477]
500 kb to 1 Mb	93.38	65	81	[1,790]	[27,200]
1 Mb to 2 Mb	49.92	36	44	[1,377]	[15,105]
2 Mb to 5 Mb	26.12	17	23	[1,45]	[8,55]
5 Mb to 10 Mb	12.77	9	11	[1,62]	[4,26]
10 Mb to 20 Mb	7.00	6	6	[1,32]	[2,14]
20 Mb to 100 Mb	3.36	1	3	[1,13]	[1,8]

Table 2: Big population statistics. The values are measured in generations.

	Mean	Mode	Median	Range	90% Range
50 kb to 100 kb	597.75	319	493	[1,6814]	[133,1422]
100 kb to 200 kb	439.88	216	351	[1,5556]	[92,1094]
200 kb to 500 kb	242.82	114	185	[1,3506]	[49,636]
500 kb to 1 Mb	94.12	54	78	[1,1767]	[23,219]
1 Mb to 2 Mb	46.07	29	39	[1,733]	[12,103]
2 Mb to 5 Mb	22.60	13	19	[1,206]	[6,50]
5 Mb to 10 Mb	10.22	7	9	[1,78]	[3,21]
10 Mb to 20 Mb	5.54	4	5	[1,30]	[2,11]
20 Mb to 100 Mb	2.65	2	2	[1,15]	[1,6]

Table 3: Small population statistics. The values are measured in generations.

	Mean	Mode	Median	Range	90% Range
50 kb to 100 kb	153.56	165	157	[2,246]	[98,198]
100 kb to 200 kb	150.14	158	154	[1,246]	[91,197]
200 kb to 500 kb	140.83	158	145	[1,246]	[75,194]
500 Kb to 1 Mb	109.74	96	108	[1,241]	[42,182]
1 Mb to 2 Mb	58.06	39	52	[1,230]	[19,118]
2 Mb to 5 Mb	27.20	19	25	[1,162]	[9,54]
5 Mb to 10 Mb	12.69	10	12	[1,56]	[4,25]
10 Mb to 20 Mb	7.12	6	7	[1,28]	[2,14]
20 Mb to 100 Mb	3.38	2	3	[1,15]	[1,7]

Table 4: Expanding population statistics. The values are measured in generations.

	Mean	Mode	Median	Range	90% Range
50 kb to 100 kb	1062.42	25	885	[2,20276]	[275,2435]
100 kb to 200 kb	528.82	25	461	[0,10544]	[133,1156]
200 kb to 500 kb	248.14	25	220	[0,2577]	[25,569]
500 kb to 1 Mb	86.71	25	60	[0,935]	[25,237]
1 Mb to 2 Mb	39.14	25	25	[0,403]	[25,109]
2 Mb to 5 Mb	26.85	25	25	[0,252]	[17,46]
5 Mb to 10 Mb	20.80	25	25	[0,90]	[5,25]
10 Mb to 20 Mb	8.00	5	6	[0,36]	[1,25]
20 Mb to 100 Mb	2.55	2	2	[0,14]	[0,7]

Table 5: Bottleneck population statistics. The values are measured in generations.
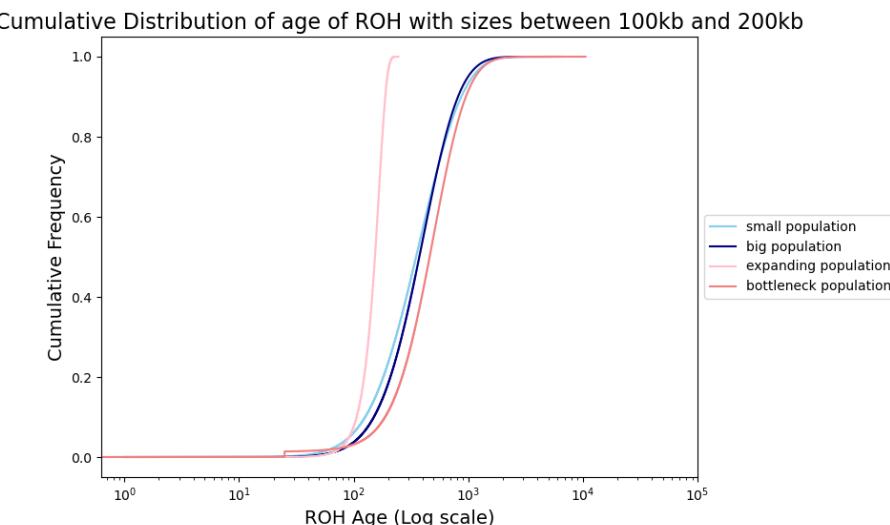
Figure 42: Cumulative plot of ROH age from segments between 100kb and 200kb comparing the four demography.
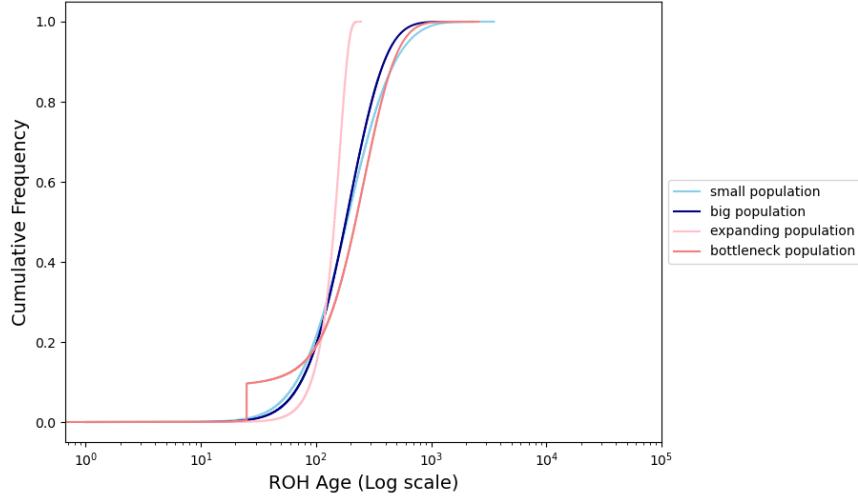
Figure 43: Cumulative plot of ROH age from segments between 200kb and 500kb comparing the four demography.
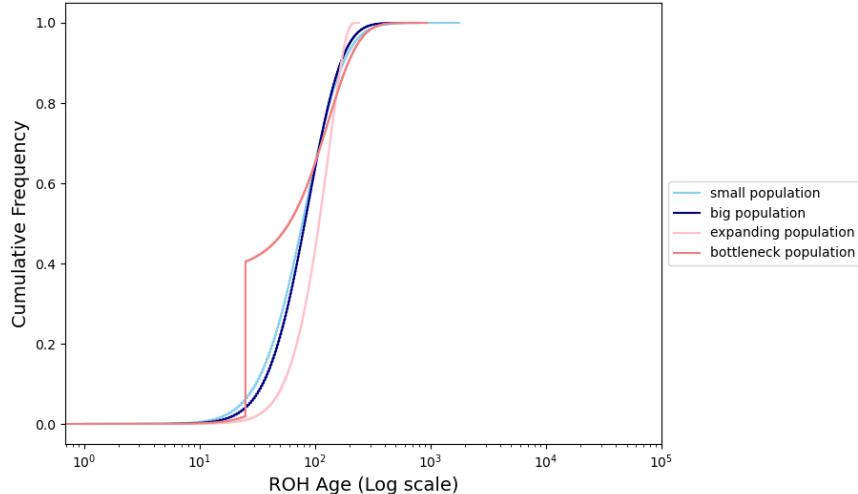

Figure 44: Cumulative plot of ROH age from segments between 500kb and 1Mb comparing the four demography.
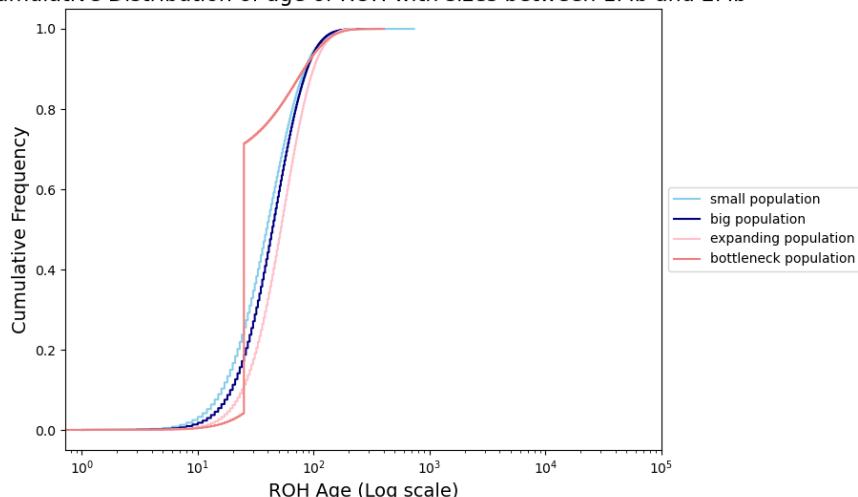

Figure 45: Cumulative plot of ROH age from segments between 1Mb and 2Mb comparing the four demography.
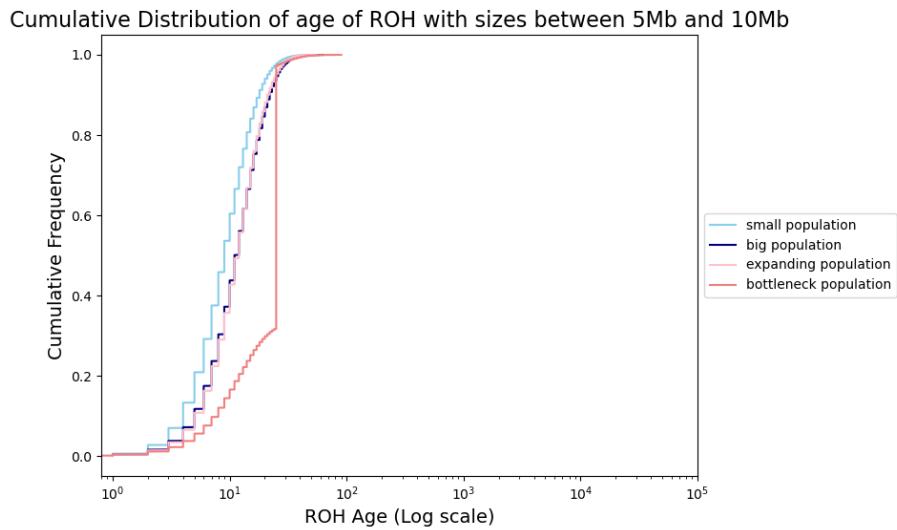
Figure 46: Cumulative plot of ROH age from segments between 5Mb and 10Mb comparing the four demography.
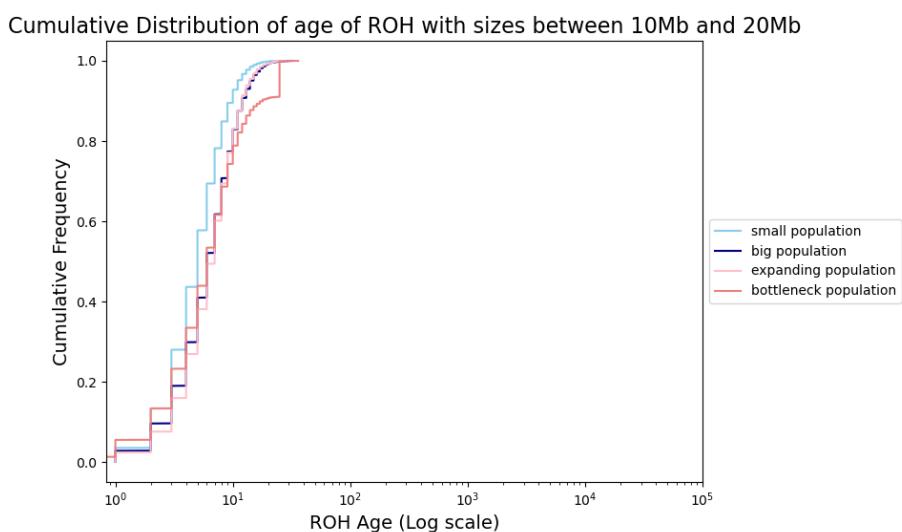

Figure 47: Cumulative plot of ROH age from segments between 10Mb and 20Mb comparing the four demography.