UNIL | Université de Lausanne
Faculté de biologie et de médecine

STATISTICAL AND BIOLOGICAL PROPERTIES OF RUNS OF HOMOZYGOSITY

Mid thesis report

Eléonore Lavanchy

PhD Program in Quantitative Biology

January 2021

Thesis director: Jérôme Goudet
Department of Ecology & Evolution,
Biophore, University of Lausanne
jerome.goudet@unil.ch

Internal expert: Roman Arguello
Department of Ecology & Evolution,
Biophore, University of Lausanne
roman.arguello@unil.ch

External expert: Daniel Wegmann
Department of Biology,
Pérolles 17, University of Fribourg
daniel.wegmann@unifr.ch

Committee president: Sven Bergmann
Department of Computational Biology
Génopode, University of Lausanne
sven.bergmann@unil.ch

Content

General Introduction

Studying relatedness between individuals is central in many areas of biology. Indeed, in quantitative genetics, it can be used to disentangle the environmental and genetic components of phenotypic variance, which has been fundamental for crops and livestock production[1–3]. Relatedness can also give insights into populations' structure[4,5] and genomic regions implicated in the determination of specific phenotypes[6–8]. Yet relatedness can only be estimated with respect to a reference because all individuals are related to some extent[9–11]. For instance, it has been shown that the most recent common ancestor in humans lived only a few thousands of years ago[12]. This reference can be the founders for pedigree data or the population itself.

Measures of relatedness rely the concept of Identical-by-descent (IBD). Genomic segments, SNPs or genes which are IBD are identical in their sequences because they are directly inherited from the same specific ancestral segment carried by one ancestor. These must be differentiated from Identical-by-State (IBS) segments which share a similar sequence but do not share a common origin – i.e. do not come from the same ancestral segment. Relatedness between individuals can be estimated through coefficients of coancestry, which are defined as the probability that two segments, genes, or SNPs drawn at random from the two individuals are IBD. Considering this, it is quite intuitive that inbreeding results in offspring harbouring large quantities of homozygous IBD segments.

Darwin was the first to show that inbreeding can be associated with various deleterious phenotypes[13,14]. This phenomenon called inbreeding depression has been suggested to play a central role in the evolution of avoidance of self-fertilization in plants[14–17] and mating between related individuals in animals[18–20]. In addition, inbreeding depression is a major concern in conservation biology. Small endangered populations harbour higher levels of relatedness due to their low effective population size resulting in a lack of genetic variability which can lead to extinction[21,22]. Two mechanisms have been proposed for explaining the deleterious effects of inbreeding. The first one is the dominance hypothesis[23] which proposes that inbreeding depression is due to the accumulation of partially to completely recessive mutations which are expressed in inbred individuals where the probability of observing homozygous genotypes is increased. The second mechanism is called the overdominance

hypothesis and states that inbreeding depression is due to heterozygous genotypes conferring higher fitness compared to homozygous genotypes. It is complicated to distinguish between both hypotheses[24,25] as physically close deleterious loci in linkage disequilibrium might cause pseudo-overdominance[17]. However, the dominance hypothesis probably applies to a larger quantity of loci compared to the overdominance hypothesis. Indeed, fewer cases are documented supporting the overdominance hypothesis[26]. In addition, lethal completely recessive mutations as well as partially recessive deleterious variants have been shown to explain most of inbreeding depression in *Drosophila*[27].

Individual levels of inbreeding can be estimated *via* inbreeding coefficients, usually denoted F and equal to the coefficient of coancestry between the parents[3,28]. Traditionally, inbreeding was measured through pedigree but with the advancement of genome sequencing technologies several genetic-based inbreeding coefficients were developed[29,30]. The majority process each SNP independently, however, there is one which rely on the proportion of homozygous IBD segments in the genome: F_{ROH}[31]. The latter is based on Runs of Homozygosity (ROHs), consecutive homozygous segments which act as a proxy for IBD segments. In addition to quantifying inbreeding and similar to IBD segments, ROHs distributions (i.e. lengths and numbers) reflect populations' past demography and history[32–34]. ROHs have also been used for identifying recessive deleterious variants with homozygosity mapping studies[35–37].

The aim of this thesis is to study biological and statistical properties of ROHs. We will first focus on understanding which types of sequencing techniques and softwares are suitable for accurate identification of ROHs. Then, we will investigate the mechanisms underlying their formation and what are their role in inbreeding and inbreeding depression. In this thesis, we plan on using mostly simulated data. However, there are genomic data for several European barn owls' populations available in the group, as well as an extensive pedigree of the swiss population which we could use to perform ROHs analyses on real data. In addition, we plan on using data available from the Human Genome Diversity Project and whitefish whole-genome-sequencing as well as Rad-sequencing data kindly provided by Claus Wedekind.

Outline

The aim of **Chapter I** is to compare different approaches developed for Runs of Homozygosity (ROHs) calling and their performances with various sequencing techniques. More precisely, we are investigating the capacity of the different methods to deal with reduced representations of the genome.

In **Chapter II,** we examine the role of Runs of Homozygosity (ROHs) in inbreeding depression. In particular, we are focussing on the distribution of different types of homozygous variants outside and inside ROHs in relation to the demographic and selective history of the population.

Chapter III focuses on investigating whether the origin i.e. the force which created a ROH between recent inbreeding, past demography or selection can be inferred. It would be based on the ROH properties, such as its length, whether the genomic region has high recombination or the proportion of individuals carrying it.

Chapter I

Introduction

Inbreeding is defined as mating between relatives and has been observed across many different taxa including humans[37–39], livestock[40–44], wild animal populations[21,45–47] and plants[21,48–50]. Its quantification as well as the understanding of its deleterious consequences – called inbreeding depression – are central in many areas of biology, from human genetics to conservation biology. Indeed, increase in genome autozygosity has been associated with many diseases in humans, such as schizophrenia[51,52] and Alzheimer's disease[53,54] as well as reduction in fertility in cattle[55,56], wolves[46] and red deer[45] and a reduction in milk productivity in cattle[55–57].

Individual levels of inbreeding are typically quantified with inbreeding coefficients (F). Traditionally, inbreeding was measured via F_{PED} through pedigrees[28]. However, this method estimates the expected inbreeding coefficient which can differ from the realized coefficient due to recombination stochasticity and mendelian segregation[58–60]. In addition, pedigree-based inbreeding calculation assumes that all founders are unrelated and non-inbred. Finally, the pedigree must be correctly recorded which is almost impossible in wild population and very difficult in domestic populations. With the advancements in sequencing technologies it became possible to estimate genomic-based inbreeding coefficients[61–64]. Many different genomic-based inbreeding coefficients have then been developed, such as F_{HOM} – implemented in the *--het* method in PLINK[29,65] – F_{UNI} and F_{GRM} (both described in[30]). All of these estimates rely on average excess SNP homozygosity and treat all SNPs independently. However, parents transmit their DNA to their offspring via large chromosomal chunks rather that each base independently. Consequently, inbreeding coefficients should be based on Identical-by-Descent (IBD) genomic segments rather than individual SNPs. Hence, a new inbreeding coefficient F_{ROH}, was proposed[31].

This novel inbreeding coefficient is based on Runs of Homozygosity (ROHs) which are long consecutive homozygous segments[37]. They arise when two Identical-by-Descent (IBD) segments are brought together in an individual as a result of parents' co-ancestry[37,66]. ROHs

were first described by Broman and Weber[66] and were shown to be ubiquitous in humans[34,37,67] and across many different taxa[47,68–70]. The ROHs-based inbreeding coefficient F_{ROH} is calculated as the proportion of the genome within ROHs and has been demonstrated to be a reliable estimator of inbreeding[31,61,63,64]. In addition, ROHs distributions (i.e. lengths and numbers) can also inform about a population's past demography and history[32–34,37,71]. Indeed, a large proportion of long ROHs reflects a high proportion of recent inbreeding while smaller ROHs indicate past relatedness among individuals in the population. Finally, ROHs can be used for identifying rare deleterious recessive variants responsible for deleterious phenotypes by homozygosity mapping[36,51,52,72,73].

Three main approaches have been developed for ROH calling: observational-based, model-based and finally likelihood-based approaches[37]. The most common method is a fast and intuitive observational-based method[37] implemented in PLINK[29,65]. It was initially developed for Whole-Genome-Sequencing (WGS) data as it assumes that the regions between two SNPs are entirely homozygous. Despite this, ROHs calling with PLINK has been applied on the output of various reduced representation sequencing techniques such as RAD-sequencing[74] or SNP arrays[44,56,71,75,76]. Given that we do not sequence the whole genome with these techniques, only a part of the SNPs is obtained and we expect that some heterozygous positions will not be sequenced and thus will be treated as homozygous by the software. Model-based approaches rely on Hidden Markov Models (HMM) and are much more computationally demanding[37]. Likelihood approaches rely on the Logarithm of Odds (LOD) of autozygosity and make use of sliding windows to estimate the ratio of the probabilities of the genotype under autozygosity and under non-autozygosity[34,66,73]. Both model and likelihood-based approaches have been outperformed by PLINK with simulated WGS data[77] but it has been suggested that model-based approaches might be better at handling sparse data[78].

The aim of this project is to use simulated data to compare ROHs calling results from WGS and reduced genomic representations to test whether these sequencing techniques are suitable for ROHs calling. We also intend on comparing the capacity of the different approaches to handle sparse data. We found that PLINK can be used to correctly estimate ROHs-based inbreeding coefficient F_{ROH} with SNP arrays and RAD-sequencing providing that a sufficient proportion of the genome has been sequenced. However, ROHs distributions

estimated with PLINK are biased with reduced genomic representations as the sum of lengths of small ROHs is constantly underestimated while the sum of lengths of large ROHs is constantly overestimated. Very preliminary results suggest that model-based approaches might be more suitable when dealing with reduced genomic data.

Material & Methods

Simulations

The simulated genomic data were produced with SLiM3, a forward-in-time individual-based simulation software[79]. The underlying model was a non-Wright Fisher model which includes overlapping generations and non-fixed effective population size[79]. The population size was regulated via a patch carrying capacity at the end of the simulation cycle and the simulation lasted for 1'000 generations. To investigate the effect of population size and genetic diversity on ROHs calling with reduced genomic data, two populations were simulated: a small population (N = 1'000) and a large population (N = 10'000). We performed ten simulation replicates for each population. For both populations, the genome architecture was composed of 30 chromosomes of 100Mb each. We decided to model a non-homogenous recombination rate along the genome. Recombination maps were then simulated with FREGENE, a forward in time simulation software[80] and based on humans recombination parameters as mentioned in Chadeau-Hyam *et al.*[80].

As the time and computational resources required to run a simulation increase linearly with the size of the simulated genomic region, each chromosome was simulated individually. To ensure that all chromosomes shared the same mating history, we ran a first batch of simulations to create a pedigree for the population. The latter was then applied to all chromosomes in the second batch of simulations. To guarantee that the simulated individuals covered the entire range of inbreeding (with coefficients from 0 to 0.5) we artificially created a small proportion of inbreeding at each generation: for each individual, the probability of mating with any individual was equal to their pedigree-based coancestry + 0.01.

As the simulations are completely neutral, the burn-in were performed afterwards, via *recapitation* in *msprime* (a python-based coalescence simulation software[81]) as suggested

by Haller *et al.*[82]. This method allows for a faster coalescence reconstruction (i.e., only for the individuals which have not coalesced yet in the forward-in-time simulation instead of a full burn-in). In addition, all mutations were added at the end of the simulation (after *recapitation*) based on a human-like mutation rate of 2.5^{e-8} [83] as suggested by Haller *et al.*[82] and Kelleher *et al.*[81].

At the end of the simulation, a random stratified sampling was performed to ensure that the individuals used for further analyses would cover the entire range of inbreeding. Individuals were selected based on pedigree inbreeding coefficients with 15 generations depth. If possible, 20 individuals were randomly selected for each 0.1 inbreeding coefficients interval from 0 to 0.5. If the interval contained less individuals, all the individuals were selected. One VCF was obtained per chromosome and then merged with *picard-tools* (http://broadinstitute.github.io/picard/) by simulation replicate.

After the individual subsampling, the final dataset consisted of ten replicates for each population size. The replicates contained 86.40 ± 4.72 individuals for the small population and 66.40 ± 3.88 individuals for the large population. The mean number of SNPs per simulation was 2'559'397.70 ± 31'990.98 SNPs for the small population and 35'409'389.20 ± 211'834.06 SNPs for the large population. Precise numbers for each simulation can be found in supplementary Table S1.

SNPs subsampling

To investigate the effect of reduced representation sequencing techniques on ROHs calling, we mimicked various reduced representation of the genome by subsampling SNPs. Figure S1 schematically describes the different subsampling methods we used. Firstly, we randomly subsampled 2%, 5%, 7%, 10%, 20%, 30%, 40%, 50%, 60%, 70% 80% and 90% of the SNPs. 100 replicates were performed for each subsampling.

To investigate the effect of filtering on rare alleles on ROHs calling, we filtered on MAF 0.01, 0.05 and 0.1 with *--maf* in VCFtools[84]. However, we are aware that allelic frequencies were biased due to the non-random subsampling of the individuals. Consequently, if we

observe an effect of MAF filtering, the same filtering should be repeated but allelic frequencies should be estimated with the entire population prior to individual subsampling.

To investigate the effect of RAD-sequencing on ROHs calling, we randomly selected 500 base-pair (bp) fragments along the genome[85] with *bedtools*[86]. Afterwards, we subsampled the SNPs that were within these fragments with the *--bed* method from *VCFtools*[84]. Because the proportion of the genome sequenced with RAD-sequencing can vary greatly depending on the organism and on which restriction enzymes were used, we varied this number of fragments. We covered between 1.3% and 33.3% of the genome for the small and between 0.03% and 3.33% for the large population. We performed 100 subsampling replicates for each proportion of genome sequenced.

To mimic SNP array sequencing, we chose two arrays initially developed for cattle and widely used for ROHs analyses: the Illumina BovineSNP 50 BeadChip (50k array) and the Illumina BovineHD BeadChip (777k array). A common feature of these two arrays is the homogenous distances between SNPs. We subsampled the SNPs for our simulated arrays based on the median distances in the existing arrays. We selected windows with size corresponding to the median distances between SNPs – 40kb for the 50k array and 3kb for the 777k array – and simply subsampled one SNP if possible (if at least one SNP was present). 100 subsampling replicates were performed for both arrays.

ROHs were identified with the *--homozyg* method, implemented in PLINK[29,65]. PLINK approach makes use of a sliding window whose size is defined by the user (*--homozyg-window-kb 5000*). For each window of X SNPs (*--homozyg-window-snp 50*), the window state (whether it is homozygous or not) is determined. The latter depends on several parameters which thresholds are defined by the user, such as the maximum number of heterozygous (*--homozyg-window-het 1*) and missing (*--homozyg-window-missing 5*) SNPs allowed in the window. Then, each SNP state is determined (whether it is within an autozygous genomic region and thus, a truly autozygous SNP or not) according to the proportion of homozygous windows, which encompassed it (*--homozyg-window-threshold 0.05*). Finally, segments are called ROHs if both the number of consecutive homozygous SNPs (*--homozyg-snp 100*) and the physical distance covered by those SNPs (*--homozyg-kb 1000*) are above certain

thresholds. Minimum SNP density in kb (*--homozyg-density 50*) and maximum distance between two adjacent SNPs (*--homozyg-gap 1000*) are also considered for a homozygous segment to be called a ROH.

For ROHs calling with PLINK, the parameters used were the same for every replicate from each subsampling method. We used the default values (mentioned in the paragraph above) for all parameters except one. We authorized zero heterozygous SNPs per windows (*--homozyg-window-het = 0*) as we used simulated data and are sure of every variant call. The minimum size required for a homozygous segment to be call a ROH was 1Mb.

Analyses

ROHs-based inbreeding coefficient F_{ROH} was estimated for each individual and is simply defined as the proportion of the genome within ROHs: $F_{ROH} = \frac{\sum Length_{ROH}}{genome\ length}$ [31]. This inbreeding coefficient was compared between WGS and all the other reduced representation sequencing technologies with Pearson correlations and linear regression. Subsampled inbreeding coefficients per individual were calculated as the mean among all subsampling replicates.

We also briefly investigated the behaviour of another inbreeding coefficient: F_{HOM} which is based on the expected under Hardy-Weinberg versus observed homozygous SNPs proportion and which is implemented in the *--het* method from PLINK[29,65]. The individual subsampled inbreeding coefficient, was estimated as the mean among subsampling replicates.

We divided ROHs into six length classes to compare the distributions between WGS and the other sequencing techniques. The length classes were: i) between 1Mb and 2Mb, ii) between 2Mb and 4Mb, iii) between 4Mb and 6Mb, iv) between 6Mb and 8Mb, v) between 8Mb and 16Mb and finally vi) larger than 16Mb. Likewise, to individual inbreeding coefficient ROHs distributions are represented as the mean sum of lengths per individual among simulation and subsampling replicates.

We calculated SNP densities with VCFtools[84] method: --*SNPdensity* for windows of 1Mb. We estimated the mean SNP density of each replicate as the mean density among the windows.

Preliminary Results

Random subsampling

All supplementary tables containing r^2, slope and intercept values per simulation replicate per subsampling method are available: https://github.com/EleonoreLavanchy/Mid-thesis-Defense.git. Figure 1 shows that, as expected, the correlation between WGS and subsampled F_{ROH} estimates increased with the proportion of SNP subsampled (Fig. 1A, Fig. S2, Table 1). However, the results differ based on population size (Fig. 1A, Fig. S2, Table 1). Indeed, in the large population, the correlation was almost perfect for all percentages (Fig. 1A, Fig. S2, Table 1). We also observed a higher variance in correlation in the small population (Fig. 1A, Fig. S2, Table 1, Table S2). Nevertheless, even if the correlation between both F_{ROH} estimates was high, the intercept was still different from zero and the slope different from 1 for smaller percentages, especially in the small population. These indicate an overestimation of F_{ROH} which slightly increases with the inbreeding coefficient of individuals with a reduced number of SNP (Fig. 1A, Fig. S2, Table S2). This intercept came close to zero and the slope close to one with 60% of the SNP in the small population and 30% of the SNP in the large population (Table 1). With two percent of the SNP, the correlation and slopes between both F_{ROH} were very low for the small population (Table 1) as almost all individual subsampled F_{ROH} estimates were equal to zero. From 5% of the SNP, the correlation between WGS and subsampled F_{ROH} was strong for the small population as well (Table 1). The relationship between both F_{HOM} is close to perfect for all percentages and both populations (Fig. S3).

Figure 1B shows that the sum of small ROHs was underestimated and the sum of large ROHs was overestimated with a reduced number of SNP for both populations (Fig. 1B, Fig. S4). As expected, and similarly to the inbreeding coefficient F_{ROH}, the results with a subsample of SNP were closer to WGS in the large population and when the proportion of subsampled SNP increased (Fig. 1B, Fig. S4). However, the distribution was still not similar to

the one of WGS even when 60% of the SNP were used (Fig. 1B). Almost no ROHs from any length class were called with 2% of the SNP in the small population (Fig. 1B). In addition, no small ROHs (1 Mb – 2Mb) were detected with 5% of the SNP in this same population and the sum of large ROHs (< 8Mb) were largely overestimated with 5% and 30% of the SNP (Fig. 1B). Concerning the large population, we observed a higher number of small (1Mb – 6Mb) than large (> 6Mb) ROHs for WGS data (Fig. 1B). However, we observed the inverse pattern (higher number of large than small ROHs) with any reduced number of SNP (Fig. 1B). The individual sum of length of large ROHs (8Mb – 16Mb) especially was overestimated with a smaller percentage of SNP (Fig. 1B).

MAF filtering

Figure 2A shows that an increase of the MAF threshold results in a departure from the values obtained with WGS (Table 2, Fig. 2A). In the large population, F_{ROH} estimates were closer to WGS when filtered on MAF = 0.01. We observed no difference between estimates based on data filtered using MAF = 0.05 and MAF = 0.1 (Table 2, Fig. 2A). Finally, the overestimation and the variance among replicates were always higher in the small compare to the large population (Table 2, Fig. 2A).

Figure 2B shows that the mean sum of small ROHs (< 6Mb) per individual was always underestimated with filtered datasets compared to the WGS-like dataset in the small population. The mean sum of large ROHs (> 8Mb) per individual was overestimated for all filtered datasets specially for MAF = 0.05 and 0.1 (Fig. 2B). We observed the same trend for the large population but the biases were of smaller magnitude (Fig. 2B).

RAD-sequencing

We identified a transitional zone in the small population where the correlation between RAD-sequencing and WGS F_{ROH} estimates goes from almost zero to one using 2.00% to 2.66% of the genome (Table 3, Fig. S5). Afterwards, when the proportion of genome sequenced increases, the inbreeding coefficients estimates get slightly overestimated, similar to what we observed with more than 5% of SNPs (Table 3, Fig. S5). Concerning the large population, the correlations were strong even with 2% of the genome (Table 3, Fig. S5).

Consequently, and as presented in Figure 3A, we reduced the minimum proportion of genome sequenced in the large population to 0.1% to match the pattern observed in the small population (Fig. 3A, Table 3). For both populations, the differences in the proportions of genome sequenced in the transitional zone are narrow, however the differences in correlation between both F_{ROH} estimates are important (Fig. 3A). As before, we observed a higher variance in correlation between both F_{ROH} estimates in the small population (especially visible when 2.2% and 2.4% of the genome are sequenced) (Table 3, Fig. 3A). As mentioned above, for the same proportion of genome sequenced, the results differed between both populations (Table 3, Fig. 3A). Consequently, we decided to investigate the effect of SNP density on the correlation between F_{ROH} estimates obtained with WGS and RAD-sequencing, presented in figure 3B. SNP density curves differ between both populations (Fig. 3B). However, the correlations reach a plateau at one for both populations for SNPs densities above two SNPs/Mb (Fig. 3B).

The sum of lengths of small (< 6Mb) ROHs was underestimated with reduced data (Fig. 3C). From the point where the correlation between WGS and subsampled F_{ROH} becomes strong, the sum of length or large ROHs is overestimated (Fig. 3C).

SNPs arrays

Figure 4A shows a strong correlation between both estimates for the small SNP array (50k) (Table 4, Fig. 5). However, with the large SNP array (777k) even if the r^2 was high, the intercept and the slope were biased compared to the small array, leading to a constant overestimation of the inbreeding coefficient (Table 4, Fig. 4A).

Concerning ROHs distributions, no small ROHs (< 6Mb) were detected with the small array and the total sum of large ROHs (> 8Mb) per individual was overestimated (Fig. 4B). For the large array, we observed the same overestimation trend for large fragments (> 8Mb). However, some small ROHs (< 6Mb) were detected (Fig. 4B).

Preliminary Discussion

ROHs calling with PLINK was initially developed for WGS data. However, it has been used on the output of various reduced representation sequencing techniques. This project focuses on investigating which types of data are suitable for ROHs analyses with PLINK. We used simulations and compared ROHs calling results between WGS and RAD-sequencing as well as SNP array like subsampling. We found that F_{ROH} can be correctly estimated, providing you have a sufficient SNP density. This minimum SNP density depends on the population genetic diversity. In addition, we found that even when the SNP density is high and leads to a high concordance between WGS and subsampled F_{ROH}, the ROHs distributions are almost always biased. Indeed, the sum of lengths of small ROHs is constantly underestimated and the sum of lengths of large ROHs is constantly overestimated.

We show that the minimum proportion of SNPs or genome sequenced required to obtain correct estimation of F_{ROH} was different between both population sizes. This could be explained by the higher genetic diversity and thus a higher number of SNPs in the large population. However, SNP density was not sufficient for predicting the relationship between both estimates, as the minimum SNP density required to obtain a correlation of one between WGS and subsampled F_{ROH} was different between both populations. It suggests that density alone is not a good proxy for determining whether the data are suitable for F_{ROH} estimation and that lower numbers of SNPs are needed to obtain reliable estimates of F_{ROH} for populations with higher genetic diversity (Fig. 3B).

Considering F_{ROH} estimations between WGS and filtered datasets, we showed that for all sequencing techniques F_{ROH} is equal to zero if the proportion of genome sequenced is too low. This is what we observed with 2% of the SNPs and 0.33% of the genome sequenced in the small population. This is due to the *--homoyg-gap* parameter from PLINK which prevents ROHs calling if the distance between two adjacent SNPs is too large resulting in no ROH in these regions. The aim is to avoid false ROHs calling in regions where no SNP were present due to bad sequencing quality. When the proportion of SNPs or genome sequenced increases, we see that F_{ROH} gets correctly estimated. This is what we observe with RAD-sequencing with 2.53% of the genome sequenced and with the small array (50k) in the small

population. To our knowledge, only one study compared ROHs calling between WGS and one SNP array – the 50k array – in cattle[87]. Their results confirm ours suggesting that individual F_{ROH} are correctly estimated with the 50k array, however they did not look at ROHs distributions. The inbreeding coefficient gets slightly overestimated when the proportion of SNPs or genome sequenced increases again as we can see with 5% of the SNPs, 8.33% of the genome sequenced and the large array (777k) in the small population. The inflation of the estimation diminishes when the proportion of genome sequenced increases further to reach results similar to WGS data as we observed with 60% of the SNPs in the small population. However reduced representation sequencing approaches rarely sequence a proportion of the genome which leads to obtaining 60% of the SNPs, except some dense SNPs arrays.

Considering MAF sequencing, the results indicated slightly higher inflation of F_{ROH} when the filtered MAF was smaller (Fig. 2A). It suggest that MAF filtering may not be used when performing ROHs analyses with PLINK which is not the case for many studies[47,88,89]. This is expected as filtering on higher values of MAF removes more variants which information are important for ROH calling. Despite this, in the small population, filtering on MAF 0.01 removed around 50% of the SNPs but the correlation, slope and intercepts were better with 50% of randomly selected SNPs (Fig. 1A) compared to results filtered on MAF 0.01 (Fig. 2A). This suggests that rare variants are important for ROHs calling. This could also be a problem for SNP arrays, as most focus on common variants, excluding are variants. Our results are supported by Meyermans *et al.*[90] who showed that it is better to avoid MAF filtering when performing ROHs analyses with PLINK. Likewise, it has been previously suggested that MAF filtering should be avoided when performing kinship estimations[9].

Regarding ROHs distributions we observed the same pattern for all reduced sequencing techniques: an underestimation of the sum of lengths of small ROHs and an overestimation of the sum of lengths of large ROHs. This bias decrease when the proportion of SNPs or genome sequenced increased. In addition, it seems that increasing the MAF filtering affects mostly large (> 8Mb) ROHs. This suggests that these rare variants are separating small adjacent ROHs in WGS data and that these small ROHs are merged into larger ROHs when MAF filtering is performed but this needs to be tested formally. It is reasonable

that these rare variants, which are probably recent mutations, are within long haplotypes where recombination had less time to act.

Our hypothesis is that, reduced representations of the genome (or reduced number of SNPs) tend to merge small adjacent ROHs into larger ones. This is supported by Zhang *et al.*[87] who mention that very long homozygous regions identified with the small SNP array in cattle are expected to be artefacts of smaller merged regions and that higher density of markers should be required to accurately identify homozygous regions[87,91]. The result of this fusion is a bias in the distributions and a slight inflation of the individual inbreeding coefficients. We hypothesize that this inflation is due to the inclusion of these small supposedly non ROHs regions between small adjacent ROHs in the total sum of ROHs and that it disappears when all SNPs between these small ROHs are sequenced. However, this needs to be tested formally. To achieve this, we plan on calculating the true and false positive and negative rates of ROHs calling on a per SNP basis as well as a visual representation of every ROHs calling in every individual. This should allow us to see precisely what is happening when ROHs are called with reduced genomic representations. This inflation may disappear with model-approaches. It is also important to note that we did not try to adapt ROHs calling parameters to our sequencing data. This could improve the results we obtained[92]. However, varying parameters should be done with caution as modifying the parameters can increase the number of ROHs called but these might not be correctly called and even individuals inbreeding ranking might get biased[90] (and results not shown). Consequently, we invite researchers to be cautious when dealing with RAD-sequencing data or SNPs arrays and ROHs analyses.

To summarize, we demonstrate that reduced genomic representations can lead to biased F_{ROH} estimates and biased ROHs distributions. However, the rank of individuals is always conserved when F_{ROH} is estimated suggesting that relative results can be trusted. If the aim is to compare individuals, reduced genomic representation can be used for ROHs calling. While we did not formally test it with ROHs distributions, total sums of lengths are always higher in the small compared to the large population and both populations distributions are similarly biased. This suggests that ROHs distributions like inbreeding coefficient could be used with reduced genomic representation for comparative studies.

However, if the aim is to precisely identify ROHs hotspots or coldspots or infer how old a ROH is, WGS data may be needed.

We only used simulated data so far and confirmation of our results with real data would be a plus. Also, the population we simulated with a small proportion of inbreeding created at each generation is not representative of natural populations, but we did it to ensure that the individuals in the population would cover the whole spectrum of inbreeding. We did not include any demography nor selection in our simulations, which is not representative of natural populations either. More realistic simulations might be needed to confirm our results. In addition, there are many different sequencing technologies for reduced genomic representation and we cannot cover them all in this manuscript. Moreover, further analyses are needed to understand precisely the consequences of ROH calling with reduced genomic data. As mentioned above other methods have been developed for ROHs calling[78,92,93]. Testing these methods on reduced genomic data and comparing the results with the ones from PLINK and with WGS data is crucial to potentially identify methods adequate for ROHs calling with sparse data.

To our knowledge, we are the first to compare F_{ROH} estimates and ROHs distributions between WGS and reduced representations of the genome. This is crucial because many studies have been performed on reduced genomic data without the insurance that the results are correct. We hope that this project will help scientists who want to perform ROHs analyses to be more careful and aware of the data they use and the analyses they carry out with these.

Further analyses

This project is a work in progress and further analyses are needed to fully understand the mechanisms underlying ROHs calling with reduced representations of the genome. First of all, we want to understand why the F_{ROH} estimates go from zero to correctly estimated and then to slightly overestimated when the proportion of genome sequenced (or SNPs subsampled) increases. We hypothesize that when F_{ROH} is correctly estimated, small adjacent ROHs are merged into larger ones with reduced genomic data. Therefore, the total sum of ROHs is correct. Small isolated ROHs have not been detected but the total sum of lengths is

compensated by the regions in between the small adjacent ROHs which are wrongly detected as homozygous. However, as mentioned above this hypothesis needs to be formally tested with true/false positive/negative of ROH calling for each SNP for each subsampling technique. Secondly, we are interested in seeing why are the ROHs distributions biased and if they are similarly biased for inbred and non-inbred individuals. To investigate both of these points, we plan on calculating true/false positive/negative rates for ROHs calling for each SNP in the simulation. This would allow us to precisely know which SNP have been assigned to a ROH (as well as the length of this ROH) for each subsampling replicate and if these regions are truly homozygous with WGS data. In addition, we plan on including visual representations presenting the autozygosity along the genome for each individual. Afterwards, every ROH that has been called could then be plotted on the corresponding genomic region. With this, we could have a precise visualization of where and which ROHs have been called for each replicate and compare it to WGS data. We already started true / false positive/ negative rates estimation as well as heterozygosity calculation along the genome but the results cannot be presented yet.

Concerning the reduced genomic technologies, we are planning on including both SNP arrays subsampling for the large population as well. We expect the F_{ROH} estimates to be better for both arrays in this second population due to the increased genetic diversity in this population. SNP selection has already been performed but we did not start the subsampling and results comparison yet.

Simulations

The simulations we performed were based on human parameters but the inbreeding we created is non-natural (never seen in nature). We are planning to build our next simulation on a real population history to have realistic distributions of inbreeding coefficients and ROHs. We decided to simulate a cattle population as ROHs analyses are extremely common for livestock. We obtained a real pedigree for a population of Belgian Blue Cattle and are planning to run one additional batch of simulations based on this pedigree. As the most common sequencing technique is SNP Array for cattle, we will compare ROHs calling results between WGS and two SNP Arrays for this simulation.

ROHs calling methods

As mentioned in the introduction, several approaches have been developed for ROHs calling. We plan on trying other methods that the observational-based PLINK method, such as the model-based *RZooRoH* method[78] or the LOD-based method implemented in *GARLIC*[93]. We already performed a pilot study for *RZooRoH* with a FirstStep student but we only used three simulations replicates from the small population. Nevertheless, it seems that *RZooRoH,* on the contrary to PLINK, can reliably estimate F_{ROH} even with 0.33% of the genome sequenced for the small population suggesting that it is indeed better at handling sparse data.

Real data

In addition to simulations, we plan on comparing the results of ROHs calling with WGS and RAD-sequencing on real data. These data consist of 37 whitefish (*Coregonus suidteri*) individuals which have been sequencing with both sequencing methods by the group of Claus Wedekind. He has kindly agreed to share these data with us. His group is currently performing variants calling and filtering and will give me the data as soon as filtering is finished. There is also WGS data for many populations of barn owls (*Tyto alba*) in Europe available in our group. We could subsample SNPs to mimic RAD-sequencing based on the restriction enzymes sites in the barn owl genome and compare ROHs calling with these data to WGS data.

Chapter II

Introduction

Mating between relatives (inbreeding) has been linked with various deleterious phenotypes in humans[38,88,94], animals[21,45,63] and plants[21,50]. This reduction of fitness in inbred individuals is called inbreeding depression and is mostly due to the accumulation of (possibly partially) recessive deleterious mutations brought to homozygous form in inbred individuals[23]. Each mutation has a specific impact on fitness and the characterization of the distribution of fitness effects (DFE) is fundamental in genetics. Despite the difference in DFE between species and even populations, empirical data suggest that specific trends are conserved, such as the exponential shape of the DFE of positively selected mutations[95–98].

Exact quantification of inbreeding is extremely difficult; however it can be approximated through the proportion of autozygosity in one individual's genome. Autozygosity itself can be approximated via Runs of Homozygosity (ROHs), long consecutive homozygous segments presumably consisting of two identical-by-descent (IBD) haplotypes brought to homozygous form by inbreeding[37,66]. The length of a ROH informs us about the time of the inbreeding event which created it[37]. Within an individual, a very long ROH indicates recent inbreeding, thus close relatedness between the parents, while a small ROH suggests ancient relatedness between the parents. In addition, ROHs have been associated with known selected genomic regions in humans[34,99] and cattle[40,44,100], suggesting that they might also appear due to selection. Indeed selection on one allele can lead to the fixation of the haplotype which contains it. This is described as the hitchhiking effect, where the increase in frequency of a beneficial allele can also impact its neighbors as they also benefit from the selection on the haplotype[101,102].

As ROHs are IBD segments, the probability of any variant to be homozygous passes from its squared allelic frequency (p^2) in genomic regions outside ROHs, to its allelic frequency (p) within ROHs[103]. The relative difference between p and p^2 is larger for variants with smaller allelic frequencies. Consequently, assuming that rare variants are more likely to be deleterious than common variants, we expect ROHs to carry more homozygous deleterious mutations than the rest of the genome. This was first proposed and demonstrated by Szpiech

et al.[103,104] in humans who showed that ROHs are enriched in homozygous deleterious variants compared to homozygous non-deleterious variants. This was then supported by the study from Zhang *et al.*[105] which showed similar results in cattle. However, the results from both studies differ in regards to ROHs enrichment in homozygous deleterious variation per ROH lengths classes. Indeed, Szpiech *et al.*[103] found that long ROHs are more enriched in homozygous deleterious variants compared small and medium ROHs in humans[103,104]. This suggests that recent inbreeding, which is responsible for long ROHs, is a major cause of inbreeding depression in humans. On the other hand, Zhang et al.[105] found that the fraction of homozygous deleterious variants is higher for small and medium ROHs compared to long ROHs in cattle, suggesting that ancient inbreeding contribute more to inbreeding depression in cattle. The authors suggest that this difference in enrichment in ROHs of different length classes is due to the population's different histories[105]. Indeed, unlike humans, cattle underwent both strong bottlenecks and broad selective pressure during breed formation[106].

In this project, we intend to use simulated data to disentangle the effect of demography and selection on the distribution of homozygous deleterious variants in the genome, with a particular focus on whether they are inside or outside ROHs, and if they are, within which ROH length class. In addition, we are planning on characterizing the distribution of damaging mutations according to their deleteriousness (i.e. selection coefficient) which both Szpiech *et al.*[103,104] and Zhang *et al.*[105] could not do with empirical data. A negative relationship has been observed with empirical data between the dominance and selective coefficient of a mutation. It is expected that both selection and dominance coefficients are to impact variants' role in inbreeding depression and their persistence in the population.

Preliminary Material & Methods

Simulations

We plan on simulating four populations with SLIM[79] each with similar size. All four populations will come from the same burn-in. The purpose of this is to ensure that the differences in the distributions of the variants derive from the demographic and selective history of the populations. The first one will be a stable (i.e. no bottleneck) and neutral population. The second one will also be a neutral population but which underwent a

bottleneck when it was separated from the ancestral population. The third one will be a stable (no bottleneck) population but where there is some selection acting. Finally, the fourth population will be the result of both a founder bottleneck and selection (scheme in Fig. S7).

We will model different variant types: (i) neutral, (ii) deleterious and (iii) positively selected. We plan on varying the dominance and selection coefficients for both deleterious and positively selected variants to precisely characterize their distributions according to their deleterious effect or positive impact. We plan on drawing selection coefficients from effect-specific distributions. Indeed, as mentioned in the introduction, it has been demonstrated that the DFE from positively selected variants tends to follow an exponential distribution[95–98]. On the other hand, it seems that the DFE of deleterious mutations might be more complicated tending towards multimodal distributions probably because different categories of deleterious mutations display different DFE[95]. DFE from mildly deleterious mutations, for instance, have been shown to tend to gamma-shaped distributions with shape parameters much inferior to one[107–109]. In addition, we intend to model variants according to the empirical inverse relationship between the dominance coefficient *h* and the selection coefficient *s*. Precise relation has been described by Henn et al.[110] such as: $h(s) = \frac{1/2}{(1+7071.07 \times s)}$ based on the work from Agrawal and Whitlock[111]. However, Kyriazis *et al.*[112] mentioned that using this equation to model large populations in SLIM is extremely time and computationally demanding. Thus, we might separate deleterious variants in three classes: slightly deleterious with $s > -0.001$ and $h = 0.5$, mildly deleterious with $-0.001 < s > -0.01$ and $h = 0.25$ and highly deleterious with $s < -0.01$ and $h = 0$ similar to what they did for large populations.

Analyses

We will first test whether our results in term of ROHs enrichment in neutral and deleterious variants are concordant with the ones from Szpiech *et al.*[103,104] and Zhang *et al.*[105], i.e. if ROHs are enriched in homozygous mutations compared to the rest of the genome. Afterwards, to disentangle the effects of demography and selection on variants distribution in the genome and assess the role of the different ROHs length classes in inbreeding depression, we plan on comparing the proportion of each type of mutations (total numbers as well as proportion in homozygous form) among the four populations within and outside ROHs and among ROHs length classes.

Chapter III

Introduction

Identification of Identical-by-Descent (IBD) genomic regions between individuals have been of interest since decades. It can be used to associate genomic regions with specific phenotypes[6–8], study relatedness between individuals [113–115], infer recent demographic history[116–118], identify regions under selection[119–121] or estimate populations' genetic parameters such as recombination[122] and mutation rates[123–125].

IBD segments can also be investigated within individuals, where they appear as consecutive homozygous segments usually called Runs of Homozygosity (ROHs) and reflect individual levels of autozygosity [31,34,37,66]. However, on the contrary to IBD segments between individuals, ROHs are entirely homozygous and thus do not require phased data, which can be a major advantage. Similarly to IBD segments, ROHs distribution between individuals (i.e. lengths and numbers) may reflect the history of the population or the individual. Long ROHs are the result of recent events of consanguineous mating and short ROHs indicate ancient inbreeding due to reduction of effective population size as result of ancient bottlenecks[32,33,37,99]. In addition, strong selection can also create regions with reduced genetic diversity. For instance, in presence of a selective sweep, the genomic variants physically close to the positively selected allele, will also reach fast fixation in a phenomenon called genetic hitchhiking[126–128]. This will result in an increase in frequency of the selected haplotype in the population. These genomic regions enriched in ROHs are named ROHs islands and have been associated with genes related to milk production in cattle[40,129] and to the gene coding for the lactase enzyme in European human populations[99].

In this chapter, we intend to investigate whether we can infer which mechanism is responsible for one particular ROH creation. For this purpose, several criteria can be used, however, to our knowledge no study tried to formally discriminate between ROHs from different origin. A first criteria which can be used is ROHs size. A long fragment spanning several mega-bases is unlikely to be the result of ancient relatedness or selection. Thompson[10] in her review suggested that the length of an IBD segment can trace back the number of generations until the most recent common ancestor. However, Speed and

Balding[130] showed that the variance associated to this measure is very large. A second criteria which could be used is the location of a ROH in the recombination landscape. If a long segment is within a region of high recombination, it is unlikely that is ancient and/or neutral. Indeed, it has been shown in humans that long ROHs, resulting from recent inbreeding, do not correlate with recombination rate while smaller ROHs appearing because of ancient bottleneck tend to cluster in low-recombination regions[67]. Finally, the proportion of the population sharing the same ROH can also be used as an indicator: if shared by many individuals it is likely that it arose via demography (common bottleneck) or selection.

Preliminary Material & Methods

To investigate whether we can infer the origin of a ROH we could use the simulations from chapter II, where we will simulate recent inbreeding, demography (with a bottleneck) and selection. Indeed we will know precisely which ROH has been created by which event and we will be able to associate the origin of a particular ROH with features such as its length, the recombination rate in this region and the proportion of individuals carrying this ROH. We could also use humans' data where the genome and the regions under selection have been extensively characterized. Finally, we could also simulate precise regions such as the region coding for lactase in humans as validation and see whether ROHs are formed and what are their characteristics.

PhD timeline

Chapter	Steps	2021		2022		2023	
		Feb – Jun	Jul – Dec	Jan – Jun	Jul – Dec	Jan – Jun	Jul – Dec
1	Simulation cattle End of subsampling *RZooRoH* ROHs calling LOD ROHs calling Writing						
2	Simulations Analyses Writing						
3	Data Preparation Analyses Writing						
Thesis	Writing						

Other additional projects might be carried out, such as a potential collaboration with a future PhD Student Alexandros Topaloudis in comparing different estimators of inbreeding for inbreeding depression quantification in barn owls.

References

1. Fisher, R. A. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
2. Wright, S. Systems of mating. I. The biometric relations between parent and offspring. *Genetics* **6**, 111 (1921).
3. Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits*. vol. 1 (Sinauer Sunderland, MA, 1998).
4. Gusev, A. *et al.* The Architecture of Long-Range Haplotypes Shared within and across Populations. *Mol. Biol. Evol.* **29**, 473–486 (2012).
5. Ralph, P. & Coop, G. The Geography of Recent Genetic Ancestry across Europe. *PLoS Biol.* **11**, e1001555 (2013).
6. Houwen, R. H. J. *et al.* Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nat. Genet.* **8**, 380–386 (1994).
7. Moltke, I., Albrechtsen, A., Hansen, T. v O., Nielsen, F. C. & Nielsen, R. A method for detecting IBD regions simultaneously in multiple individuals—with applications to disease genetics. *Genome Res.* **21**, 1168–1180 (2011).
8. Browning, S. R. & Thompson, E. A. Detecting Rare Variant Associations by Identity-by-Descent Mapping in Case-Control Studies. *Genetics* **190**, 1521–1531 (2012).
9. Goudet, J., Kay, T. & Weir, B. S. How to estimate kinship. *Mol. Ecol.* **27**, 4121–4135 (2018).
10. Thompson, E. A. Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics* **194**, 301–326 (2013).
11. Wang, J. Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theor. Popul. Biol.* **107**, 4–13 (2016).
12. Rohde, D. L. T., Olson, S. & Chang, J. T. Modelling the recent common ancestry of all living humans. *Nature* **431**, 562–566 (2004).
13. Darwin, C. *The variation of animals and plants under domestication*. vol. 2 (Cambridge University Press, 1868).
14. Darwin, C. *The effects of cross and self fertilization in the vegetable kingdom*. (John Murray, 1876).
15. Darwin, C. *The different forms of flowers on plants of the same species*. (D. Appleton and Company, 1897).
16. Charlesworth, D. & Charlesworth, B. Inbreeding depression and its evolutionary consequences. *Annu. Rev. Ecol. Syst.* 237–268 (1987).
17. Hedrick, P. W., Hellsten, U. & Grattapaglia, D. Examining the cause of high inbreeding depression: analysis of whole-genome sequence data in 28 selfed progeny of Eucalyptus grandis. *New Phytol.* **209**, 600–611 (2016).
18. Shields, W. M. *Philopatry, inbreeding, and the evolution of sex*. (SUNY press, 1982).
19. Thornhill, N. W. *The natural history of inbreeding and outbreeding: theoretical and empirical perspectives*. (University of Chicago Press, 1993).
20. Pusey, A. & Wolf, M. Inbreeding avoidance in animals. *Trends Ecol. Evol.* **11**, 201–206 (1996).
21. Keller, L. F. & Waller, D. M. Inbreeding effects in wild populations. *Trends Ecol. Evol.* **17**, 230–241 (2002).
22. Hedrick, P. W. & Garcia-Dorado, A. Understanding inbreeding depression, purging, and genetic rescue. *Trends Ecol. Evol.* **31**, 940–952 (2016).
23. Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. *Nat. Rev. Genet.* **10**, 783–796 (2009).
24. Crow, J. F. Mutation, mean fitness, and genetic load. *Oxf. Surv Evol Biol* **9**, 3–42 (1993).
25. Lee, Y. W., Fishman, L., Kelly, J. K. & Willis, J. H. A Segregating Inversion Generates Fitness Variation in Yellow Monkeyflower (*Mimulus guttatus*). *Genetics* **202**, 1473–1484 (2016).
26. Hedrick, P. W. What is the evidence for heterozygote advantage selection? *Trends Ecol. Evol.* **27**, 698–704 (2012).

27. Simmons, M. J. & Crow, J. F. Mutations affecting fitness in Drosophila populations. *Annu. Rev. Genet.* **11**, 49–78 (1977).
28. Wright, S. Coefficients of Inbreeding and Relationship. *Am. Nat.* **56**, 330–338 (1922).
29. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
30. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
31. McQuillan, R. *et al.* Runs of Homozygosity in European Populations. *Am. J. Hum. Genet.* **83**, 359–372 (2008).
32. Nothnagel, M., Lu, T. T., Kayser, M. & Krawczak, M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Hum. Mol. Genet.* **19**, 2927–2935 (2010).
33. Kirin, M. *et al.* Genomic Runs of Homozygosity Record Population History and Consanguinity. *PLoS ONE* **5**, e13996 (2010).
34. Pemberton, T. J. *et al.* Genomic Patterns of Homozygosity in Worldwide Human Populations. *Am. J. Hum. Genet.* **91**, 275–292 (2012).
35. Ku, C. S., Naidoo, N., Teo, S. M. & Pawitan, Y. Regions of homozygosity and their impact on complex diseases and traits. *Hum. Genet.* **129**, 1–15 (2011).
36. Alkuraya, F. S. The application of next-generation sequencing in the autozygosity mapping of human recessive diseases. *Hum. Genet.* **132**, 1197–1211 (2013).
37. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows into population history and trait architecture. *Nat. Rev. Genet.* **19**, 220–234 (2018).
38. Bittles, A. H. & Black, M. L. Consanguinity, human evolution, and complex diseases. *Proc. Natl. Acad. Sci.* **107**, 1779–1786 (2010).
39. Leutenegger, A.-L., Sahbatou, M., Gazal, S., Cann, H. & Génin, E. Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? *Eur. J. Hum. Genet.* **19**, 583–587 (2011).
40. Kim, E.-S. *et al.* Effect of Artificial Selection on Runs of Homozygosity in U.S. Holstein Cattle. *PLoS ONE* **8**, e80813 (2013).
41. Peripolli, E. *et al.* Runs of homozygosity: current knowledge and applications in livestock. *Anim. Genet.* **48**, 255–271 (2017).
42. Peripolli, E. *et al.* Assessment of runs of homozygosity islands and estimates of genomic inbreeding in Gyr (Bos indicus) dairy cattle. *BMC Genomics* **19**, 34 (2018).
43. Howard, J. T., Pryce, J. E., Baes, C. & Maltecca, C. Invited review: Inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genomic variability. *J. Dairy Sci.* **100**, 6009–6024 (2017).
44. Forutan, M. *et al.* Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. *BMC Genomics* **19**, 98 (2018).
45. Huisman, J., Kruuk, L. E. B., Ellis, P. A., Clutton-Brock, T. & Pemberton, J. M. Inbreeding depression across the lifespan in a wild mammal population. *Proc. Natl. Acad. Sci.* **113**, 3585–3590 (2016).
46. Åkesson, M. *et al.* Genetic rescue in a severely inbred wolf population. *Mol. Ecol.* **25**, 4745–4756 (2016).
47. Kardos, M. *et al.* Genomic consequences of intensive inbreeding in an isolated wolf population. *Nat. Ecol. Evol.* **2**, 124–131 (2018).
48. Menges, E. S. Seed Germination Percentage Increases with Population Size in a Fragmented Prairie Species. *Conserv. Biol.* **5**, 158–164 (1991).
49. Kariyat, R. R. & Stephenson, A. G. Inbreeding depression: it's not just for population biologists. *Am. J. Bot.* **106**, 331–333 (2019).
50. Zhang, C. *et al.* The genetic basis of inbreeding depression in potato. *Nat. Genet.* **51**, 374–378 (2019).
51. Lencz, T. *et al.* Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl. Acad. Sci.* **104**, 19942–19947 (2007).
52. Keller, M. C. *et al.* Runs of Homozygosity Implicate Autozygosity as a Schizophrenia Risk Factor. *PLoS Genet.* **8**, e1002656 (2012).

53. Nalls, M. A. *et al.* Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *neurogenetics* **10**, 183–190 (2009).

54. Ghani, M. *et al.* Association of Long Runs of Homozygosity With Alzheimer Disease Among African American Individuals. *JAMA Neurol.* **72**, 1313–1323 (2015).

55. Parland, S. M., Kearney, J. F., Rath, M. & Berry, D. P. Inbreeding Effects on Milk Production, Calving Performance, Fertility, and Conformation in Irish Holstein-Friesians. *J. Dairy Sci.* **90**, 4411–4419 (2007).

56. Bjelland, D. W., Weigel, K. A., Vukasinovic, N. & Nkrumah, J. D. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. *J. Dairy Sci.* **96**, 4697–4706 (2013).

57. Pryce, J. E., Haile-Mariam, M., Goddard, M. E. & Hayes, B. J. Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. *Genet. Sel. Evol.* **46**, 71 (2014).

58. Franklin, I. R. The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. *Theor. Popul. Biol.* **11**, 60–80 (1977).

59. Carothers, A. D. *et al.* Estimating Human Inbreeding Coefficients: Comparison of Genealogical and Marker Heterozygosity Approaches. *Ann. Hum. Genet.* **70**, 666–676 (2006).

60. Hill, W. G. & Weir, B. S. Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genet. Res.* **93**, 47–64 (2011).

61. Keller, M. C., Visscher, P. M. & Goddard, M. E. Quantification of Inbreeding Due to Distant Ancestors and Its Detection Using Dense Single Nucleotide Polymorphism Data. *Genetics* **189**, 237–249 (2011).

62. Hoffman, J. I. *et al.* High-throughput sequencing reveals inbreeding depression in a natural population. *Proc. Natl. Acad. Sci.* **111**, 3775–3780 (2014).

63. Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G. & Allendorf, F. W. Genomics advances the study of inbreeding depression in the wild. *Evol. Appl.* **9**, 1205–1218 (2016).

64. Alemu, S. W. *et al.* An evaluation of inbreeding measures using a whole-genome sequenced cattle pedigree. *Heredity* (2020) doi:10.1038/s41437-020-00383-9.

65. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

66. Broman, K. W. & Weber, J. L. Long Homozygous Chromosomal Segments in Reference Families from the Centre d'Étude du Polymorphisme Humain. *Am. J. Hum. Genet.* **65**, 1493–1500 (1999).

67. Gibson, J., Morton, N. E. & Collins, A. Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* **15**, 789–795 (2006).

68. Ferencakovic, M., Hamzic, E., Gredler-Grandl, B., Curik, I. & Sölkner, J. Runs of Homozygosity Reveal Genome-wide Autozygosity in the Austrian Fleckvieh Cattle. *Agric. Conspec. Sci.* **76**, 325–328 (2011).

69. Saremi, N. F. *et al.* Puma genomes from North and South America provide insights into the genomic consequences of inbreeding. *Nat. Commun.* **10**, 4769 (2019).

70. Liu, L. *et al.* Genetic consequences of long-term small effective population size in the critically endangered pygmy hog. *Evol. Appl.* **00**, (2020).

71. Bosse, M. *et al.* Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape. *PLoS Genet.* **8**, e1003100 (2012).

72. Hildebrandt, F. *et al.* A Systematic Approach to Mapping Recessive Disease Genes in Individuals from Outbred Populations. *PLoS Genet.* **5**, e1000353 (2009).

73. Wang, S., Haynes, C., Barany, F. & Ott, J. Genome-wide autozygosity mapping in human populations. *Genet. Epidemiol.* **33**, 172–180 (2009).

74. Grossen, C., Biebach, I., Angelone-Alasaad, S., Keller, L. F. & Croll, D. Population genomics analyses of European ibex species show lower diversity and higher inbreeding in reintroduced populations. *Evol. Appl.* **11**, 123–139 (2018).

75. Gurgul, A. *et al.* The use of runs of homozygosity for estimation of recent inbreeding in Holstein cattle. *J. Appl. Genet.* **57**, 527–530 (2016).

76. de Jong, J. F. *et al.* Fragmentation and Translocation Distort the Genetic Landscape of Ungulates: Red Deer in the Netherlands. *Front. Ecol. Evol.* **8**, 365 (2020).

77. Howrigan, D. P., Simonson, M. A. & Keller, M. C. Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics* **12**, 460 (2011).

78. Bertrand, A. R., Kadri, N. K., Flori, L., Gautier, M. & Druet, T. RZooRoH: An R package to characterize individual genomic autozygosity and identify homozygous-by-descent segments. *Methods Ecol. Evol.* **10**, 860–866 (2019).

79. Haller, B. C. & Messer, P. W. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Mol. Biol. Evol.* **36**, 632–637 (2019).

80. Chadeau-Hyam, M. *et al.* Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics* **9**, 364 (2008).

81. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Comput. Biol.* 22 (2016).

82. Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W. & Ralph, P. L. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Mol. Ecol. Resour.* **19**, 552–566 (2019).

83. Nachman, M. W. & Crowell, S. L. Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics* **156**, 297–304 (2000).

84. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

85. Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G. & Hohenlohe, P. A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **17**, 81–92 (2016).

86. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinforma.* **47**, 11.12.1-11.12.34 (2014).

87. Zhang, Q., Calus, M. P. L., Guldbrandtsen, B., Lund, M. S. & Sahana, G. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. *BMC Genet.* **16**, (2015).

88. Clark, D. W. *et al.* Associations of autozygosity with a broad range of human phenotypes. *Nat. Commun.* **10**, 4957 (2019).

89. Ghoreishifar, S. M. *et al.* Genomic measures of inbreeding coefficients and genome-wide scan for runs of homozygosity islands in Iranian river buffalo, Bubalus bubalis. *BMC Genet.* **21**, 16 (2020).

90. Meyermans, R., Gorssen, W., Buys, N. & Janssens, S. How to study runs of homozygosity using PLINK? A guide for analyzing medium density SNP data in livestock and pet species. *BMC Genomics* **21**, 94 (2020).

91. Ramey, H. R. *et al.* Detection of selective sweeps in cattle using genome-wide SNP data. *BMC Genomics* **14**, 382 (2013).

92. Ceballos, F. C., Hazelhurst, S. & Ramsay, M. Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data. *BMC Genomics* **19**, 106 (2018).

93. Szpiech, Z. A., Blant, A. & Pemberton, T. J. GARLIC: Genomic Autozygosity Regions Likelihood-based Inference and Classification. *Bioinformatics* **33**, 2059–2062 (2017).

94. Howrigan, D. P. *et al.* Genome-wide autozygosity is associated with lower general cognitive ability. *Mol. PSYCHIATRY* **21**, 837–843 (2016).

95. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610–618 (2007).

96. Kassen, R. & Bataillon, T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nat. Genet.* **38**, 484–488 (2006).

97. Orr, H. A. The Distribution of Fitness Effects Among Beneficial Mutations. *Genetics* **163**, 1519–1526 (2003).

98. Rokyta, D. R., Joyce, P., Caudle, S. B. & Wichman, H. A. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nat. Genet.* **37**, 441–444 (2005).

99. Curtis, D., Vine, A. E. & Knight, J. Study of Regions of Extended Homozygosity Provides a Powerful Method to Explore Haplotype Structure of Human Populations. *Ann. Hum. Genet.* **72**, 261–278 (2008).

100. Peripolli, E. *et al.* Genome-wide detection of signatures of selection in indicine and Brazilian locally adapted taurine cattle breeds using whole-genome re-sequencing data. *BMC Genomics* **21**, 624 (2020).

101. McVean, G. The Structure of Linkage Disequilibrium Around a Selective Sweep. *Genetics* **175**, 1395–1406 (2007).

102. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **89**, 391–403 (2007).
103. Szpiech, Z. A. *et al.* Long Runs of Homozygosity Are Enriched for Deleterious Variation. *Am. J. Hum. Genet.* **93**, 90–102 (2013).
104. Szpiech, Z. A. *et al.* Ancestry-Dependent Enrichment of Deleterious Homozygotes in Runs of Homozygosity. *Am. J. Hum. Genet.* **105**, 747–762 (2019).
105. Zhang, Q., Guldbrandtsen, B., Bosse, M., Lund, M. S. & Sahana, G. Runs of homozygosity and distribution of functional variants in the cattle genome. *BMC Genomics* **16**, 542 (2015).
106. Frantz, L. A. F., Bradley, D. G., Larson, G. & Orlando, L. Animal domestication in the era of ancient genomics. *Nat. Rev. Genet.* **21**, 449–460 (2020).
107. Eyre-Walker, A., Woolfit, M. & Phelps, T. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. *Genetics* **173**, 891–900 (2006).
108. Loewe, L. & Charlesworth, B. Inferring the distribution of mutational effects on fitness in Drosophila. *Biol. Lett.* **2**, 426–430 (2006).
109. Loewe, L., Charlesworth, B., Bartolomé, C. & Nöel, V. Estimating Selection on Non-synonymous Mutations. *Genetics* **172**, 1079–1092 (2006).
110. Henn, B. M. *et al.* Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc. Natl. Acad. Sci.* **113**, E440–E449 (2016).
111. Agrawal, A. F. & Whitlock, M. C. Inferences About the Distribution of Dominance Drawn From Yeast Gene Knockout Data. *Genetics* **187**, 553–566 (2011).
112. Kyriazis, C. C., Wayne, R. K. & Lohmueller, K. E. Strongly deleterious mutations are a primary determinant of extinction risk due to inbreeding depression. *Evol. Lett.* **n/a**,.
113. Huff, C. D. *et al.* Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res.* **21**, 768–774 (2011).
114. Ramstetter, M. D. *et al.* Benchmarking Relatedness Inference Methods with Genome-Wide Data from Thousands of Relatives. *Genetics* **207**, 75–82 (2017).
115. Qiao, Y., Sannerud, J. G., Basu-Roy, S., Hayward, C. & Williams, A. L. Distinguishing pedigree relationships via multi-way identity by descent sharing and sex-specific genetic maps. *Am. J. Hum. Genet.* **108**, 68–83 (2021).
116. Palamara, P. F., Lencz, T., Darvasi, A. & Pe'er, I. Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. *Am. J. Hum. Genet.* **91**, 809–822 (2012).
117. Browning, S. R. & Browning, B. L. Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *Am. J. Hum. Genet.* **97**, 404–418 (2015).
118. Browning, S. R. *et al.* Ancestry-specific recent effective population size in the Americas. *PLOS Genet.* **14**, e1007385 (2018).
119. Albrechtsen, A., Moltke, I. & Nielsen, R. Natural Selection and the Distribution of Identity-by-Descent in the Human Genome. *Genetics* **186**, 295–308 (2010).
120. Han, L. & Abney, M. Using identity by descent estimation with dense genotype data to detect positive selection. *Eur. J. Hum. Genet.* **21**, 205–211 (2013).
121. Palamara, P. F., Terhorst, J., Song, Y. S. & Price, A. L. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nat. Genet.* **50**, 1311–1317 (2018).
122. Zhou, Y., Browning, S. R. & Browning, B. L. A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. *Am. J. Hum. Genet.* **106**, 426–437 (2020).
123. Campbell, C. D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44**, 1277–1281 (2012).
124. Narasimhan, V. M. *et al.* Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat. Commun.* **8**, 303 (2017).
125. Tian, X., Browning, B. L. & Browning, S. R. Estimating the Genome-wide Mutation Rate with Three-Way Identity by Descent. *Am. J. Hum. Genet.* **105**, 883–893 (2019).
126. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
127. Kaplan, N. L., Hudson, R. R. & Langley, C. H. The 'hitchhiking effect' revisited. *Genetics* **123**, 887–899 (1989).

128. Charlesworth, B., Charlesworth, D. & Barton, N. H. The Effects of Genetic and Geographic Structure on Neutral Variation. *Annu. Rev. Ecol. Evol. Syst.* **34**, 99–125 (2003).
129. Moscarelli, A. *et al.* Genome-wide assessment of diversity and differentiation between original and modern Brown cattle populations. *Anim. Genet.* **52**, 21–31 (2020).
130. Speed, D. & Balding, D. J. Relatedness in the post-genomic era: is it still useful? *Nat. Rev. Genet.* **16**, 33–44 (2015).

Appendix

FIGURES

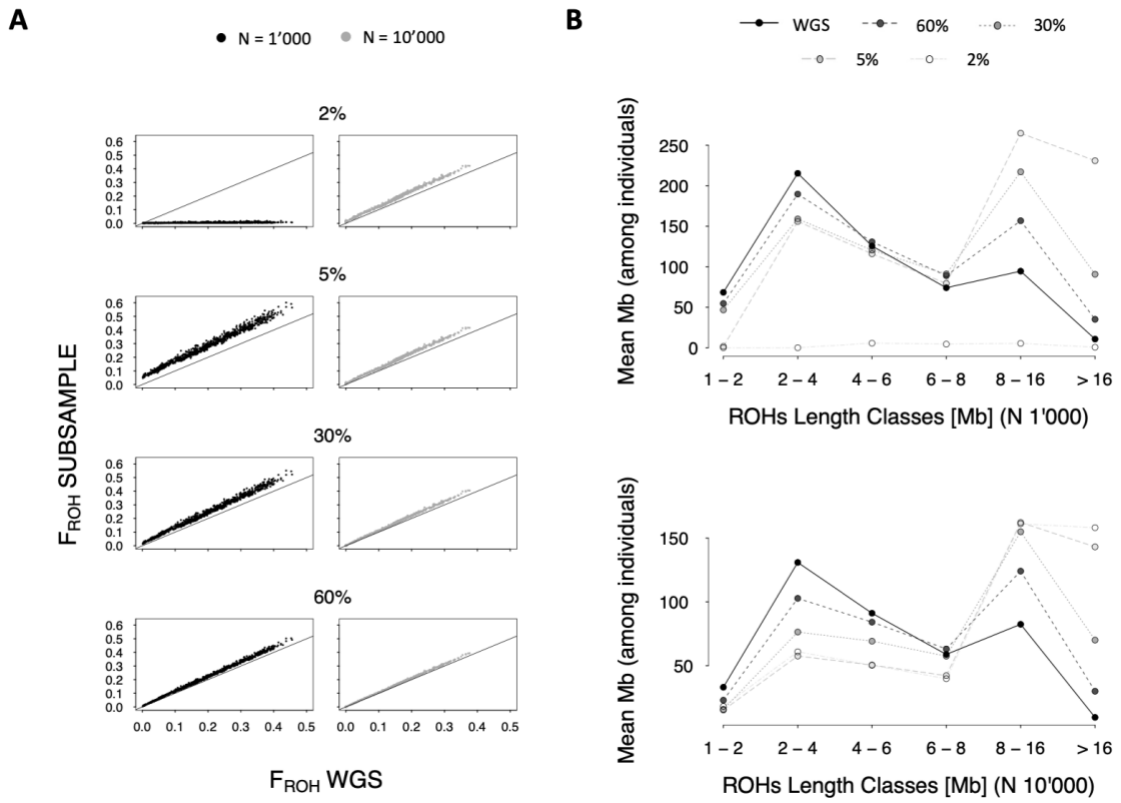

Figure 1: A: *The relationship between F_{ROH} estimated with a subsample of the SNPs (indicated in percentages above plots) in the y axis and F_{ROH} calculated with all the SNPs in the x axis for both populations (small is in back on the left and large is in grey on the right). The black lines represent the perfect correlation between both. Each point represents one individual. All simulation replicates are represented in these graphs. F_{ROH} subsampled is represented as the mean among subsampling replicates. r^2, slopes and intercepts are reported in Table 2.* ***B:*** *ROHs distributions. The mean individual sum of lengths of ROHs on the y axis per ROHs length classes on the x axis. Small population is above and large population is below. Individual means are among subsampling and simulations replicates. Points for the same subsampling are linked with specific lines. Legend for both plots on top of the first plot.*

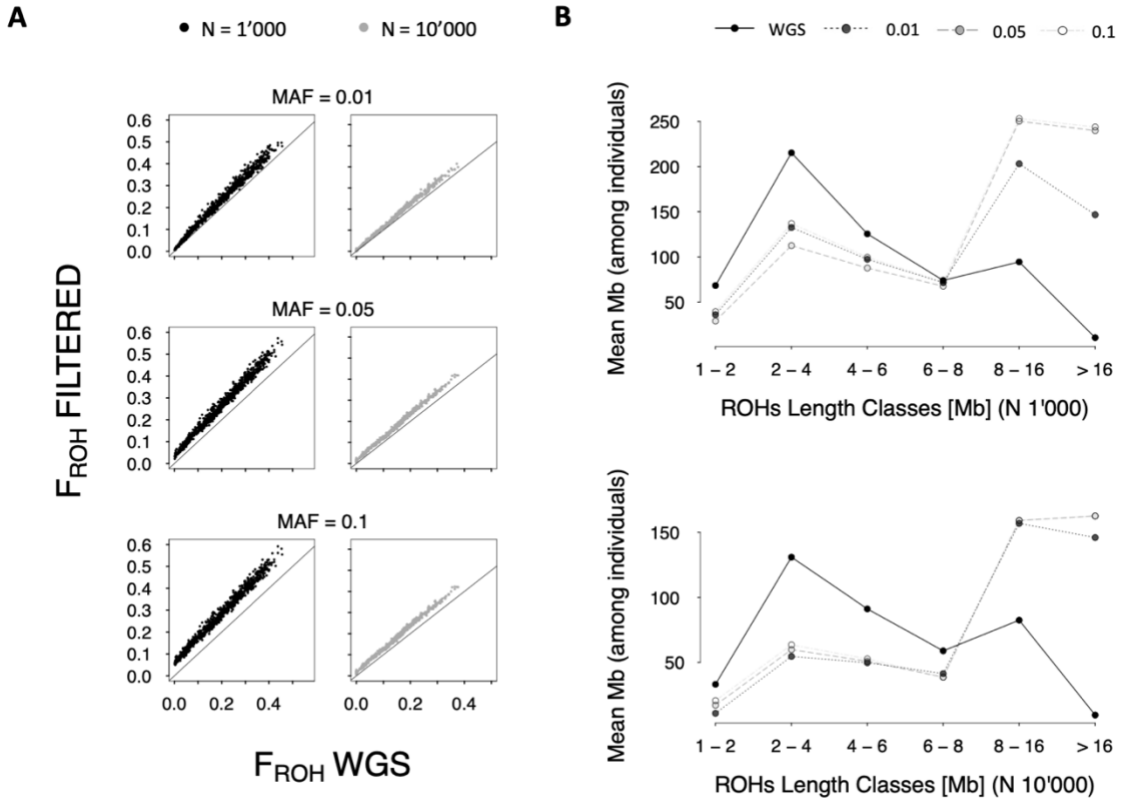

Figure 2: A: *The relationship between F_{ROH} estimated with filtering performed on MAF (MAF above plots) on the y axis and F_{ROH} calculated with no filtering on the x axis for both populations (small is in back on the left and large is in grey on the right). The black lines represent the perfect correlation between both. Each point represents one individual. All simulation replicates are represented in these graphs. r^2, slopes and intercepts are reported in Table 3.* ***B:*** *ROHs distributions. The mean individual sum of lengths of ROHs on the y axis per ROHs length classes on the x axis. Small population is above and large population is below. Individual means are among simulations replicates. Points for the same filtering are linked with specific lines. Legend for both plots on top of the first plot.*

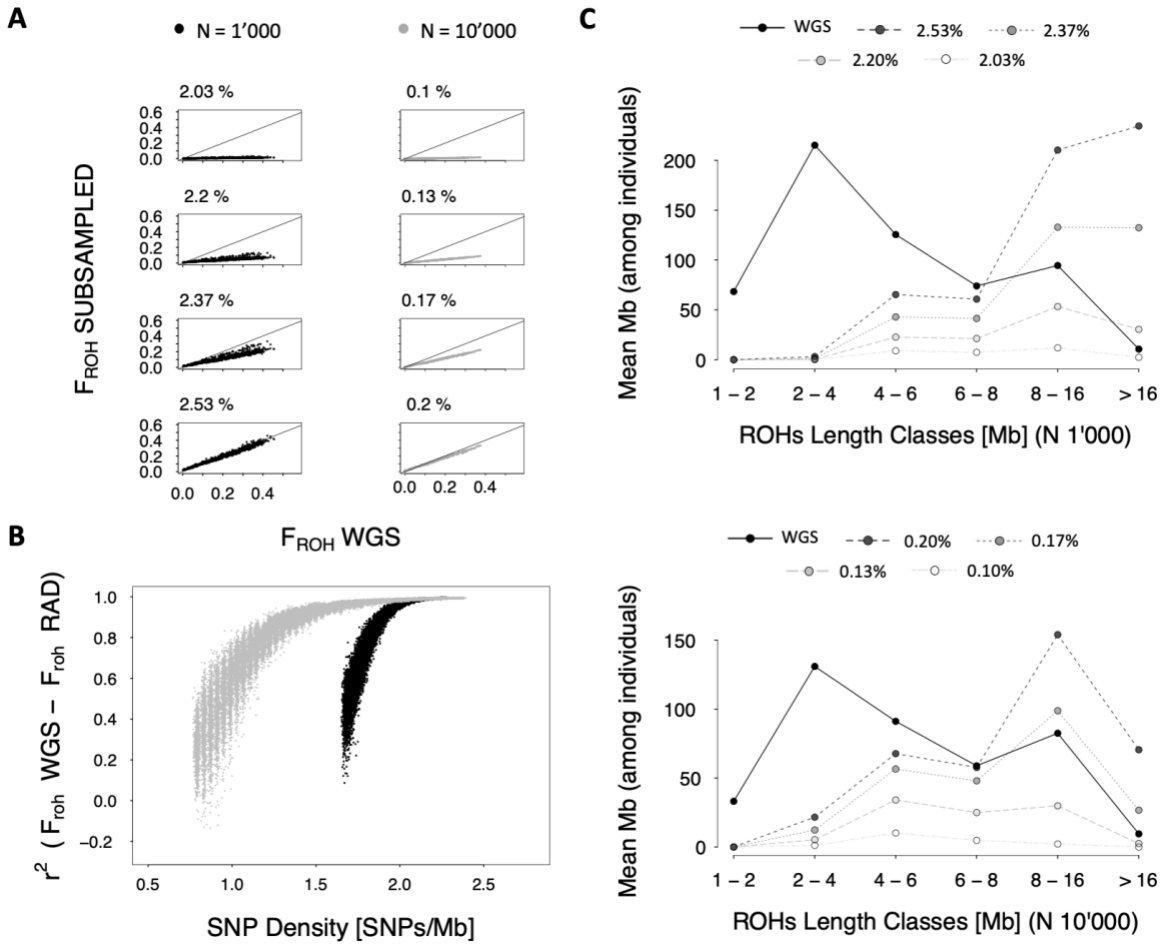

Figure 3: A: *The relationship between F_{ROH} estimated with a subsample of the genome (indicated in percentages above plots) mimicking RAD-sequencing on the y axis and F_{ROH} calculated with all the data (WGS-like) on the x axis for both populations (small is in back on the left and large is in grey on the right). The black lines represent the perfect correlation between both. Each point represents one individual. All simulation replicates are represented in these graphs. F_{ROH} RAD-sequencing is represented as the mean among subsampling replicates. r^2, slopes and intercepts are reported in Table 4.* ***B:*** *Correlation between F_{ROH} WGS (on the x axis) and F_{ROH} with RAD-sequencing-like data (on the y axis) dependent on mean SNP density for the subsampling replicate [SNPs/Mb]. All simulations are represented in this graph and each point represents the correlation for one subsampling replicate. Same colours as panel A, small population in black and large population in grey.* ***C:*** *ROHs distributions. The mean individual sum of lengths of ROHs on the y axis per ROHs length classes on the x axis. Small population is above and large population is below. Individual means are among subsampling and simulations replicates. Points for the same subsampling are linked with specific lines. Legends above each plot.*

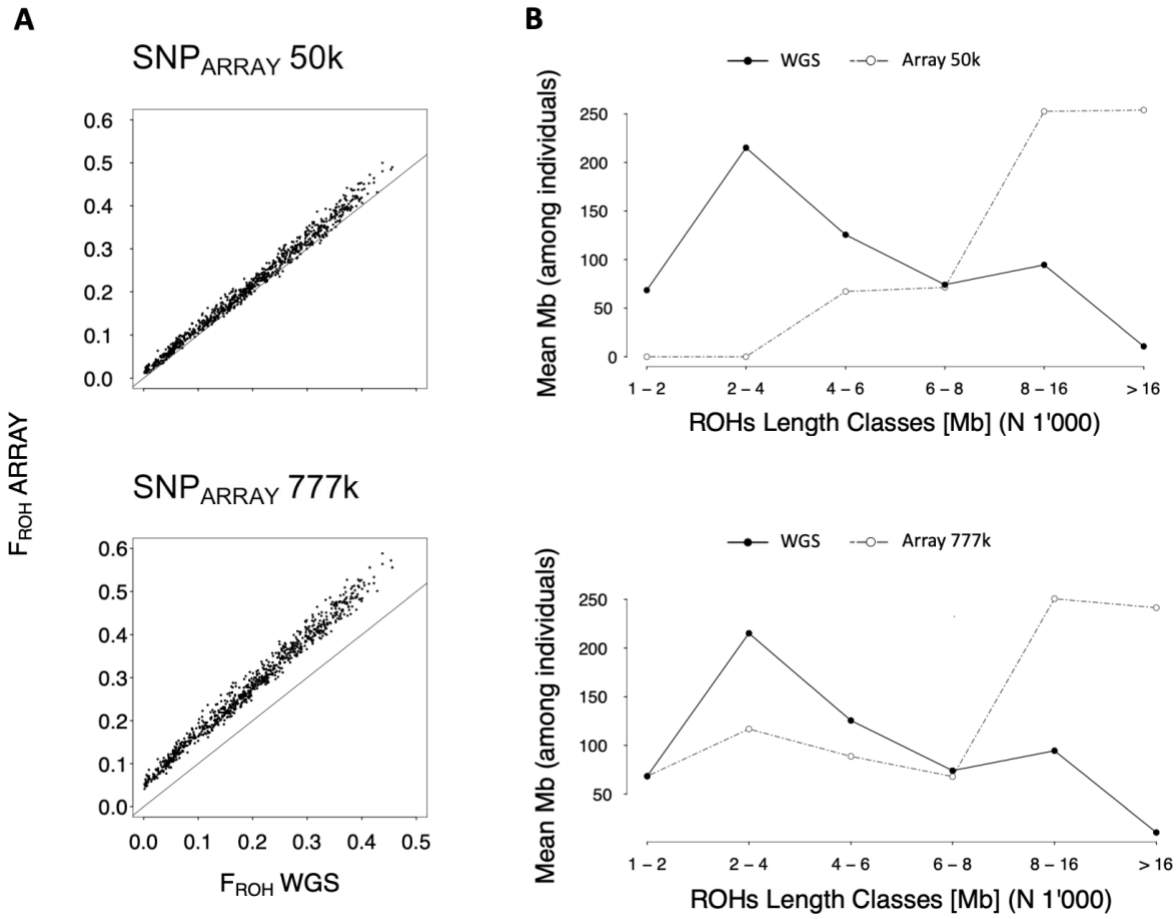

Figure 4: *The relationship between F_{ROH} estimated with SNP Arrays like subsampling on the y axis and F_{ROH} calculated with all data (WGS) on the x axis for the small population. The graph above represents SNP Array 50k subsampling and the second graph below SNP Array 700k subsampling. The black lines represent the perfect correlation between both. Each point represents one individual. All simulation replicates are represented in these graphs. Mean r^2, slopes and intercepts are reported in Table 5. r^2, slopes and intercepts per replicate are reported in Table S5. B: ROHs distributions. The mean individual sum of lengths of ROHs on the y axis per ROHs length classes on the x axis. Small SNP Array (50k) is above and large SNP Array (700k) is below. Individual means are among simulations replicates. Legends on top of the plots.*

TABLES

Table 1: *Mean ± SD r^2, slopes and intercepts per SNP subsampling percentages for both populations. Means are among subsampling and simulation replicates. Exact Mean r^2, slopes and intercepts per simulation replicate can be found in Table S2. Rows are ordered by population size first and SNPs subsampling percentage second.*

N Population	% SNPs subsampled	r^2	Slope	Intercept
1'000	2	0.152 ± 0.063	0.014 ± 0.006	0.002 ± 0.001
1'000	5	0.985 ± 0.003	1.137 ± 0.020	0.059 ± 0.003
1'000	30	0.992 ± 0.001	1.130 ± 0.150	0.020 ± 0.002
1'000	60	0.997 ± 0.000	1.080 ± 0.009	0.007 ± 0.001
10'000	2	0.996 ± 0.001	1.110 ± 0.004	0.014 ± 0.001
10'000	5	0.997 ± 0.001	1.106 ± 0.004	0.009 ± 0.001

10’000	30	0.998 ± 0.000	1.085 ± 0.004	0.003 ± 0.000
10’000	60	0.999 ± 0.000	1.052 ± 0.002	0.001 ± 0.000

Table 2: *Mean ± SD r^2, slopes and intercepts per MAF filtering for both populations. Means are among simulation replicates. Exact Mean r^2, slopes and intercepts per simulation replicate can be found in Table S3. Rows are ordered by population size first and MAF second.*

N Population	MAF	r^2	Slope	Intercept
1’000	0.01	0.989 ± 0.001	1.096 ± 0.020	0.014 ± 0.003
1’000	0.05	0.987 ± 0.003	1.138 ± 0.013	0.039 ± 0.003
1’000	0.1	0.985 ± 0.003	1.115 ± 0.017	0.062 ± 0.004
10’000	0.01	0.997 ± 0.001	1.095 ± 0.010	0.006 ± 0.001
10’000	0.05	0.996 ± 0.001	1.096 ± 0.003	0.015 ± 1.001
10’000	0.1	0.996 ± 0.001	1.092 ± 0.004	0.018 ± 0.001

Table 3: *Mean ± SD r^2, slopes and intercepts with RAD-sequencing-like subsampling for both populations. Means are among subsampling and simulation replicates. Exact Mean r^2, slopes and intercepts per simulation replicate can be found in Table S4. Rows are ordered by population size first and percentage of genome sequenced second.*

N Population	% genome sequenced	r^2	Slope	Intercept
1000	1.333	0.000 ± 0.000	0.000 ± 0.000	0.000± 0.000
1000	1.667	0.003 ± 0.002	0.000 ± 0.000	0.000 ± 0.000
1000	2.000	0.220 ± 0.061	0.022 ± 0.008	0.003 ± 0.001
1000	2.033	0.282 ± 0.076	0.032 ± 0.011	0.004 ± 0.001
1000	2.067	0.36 ± 0.078	0.046 ± 0.016	0.005 ± 0.001
1000	2.100	0.446 ± 0.084	0.065 ± 0.021	0.006 ± 0.001
1000	2.133	0.529 ± 0.078	0.091 ± 0.029	0.007 ± 0.001
1000	2.167	0.609 ± 0.074	0.125 ± 0.038	0.008 ± 0.001
1000	2.200	0.687 ± 0.067	0.170 ± 0.049	0.009 ± 0.001
1000	2.233	0.743 ± 0.058	0.223 ± 0.059	0.010 ± 0.001
1000	2.267	0.802 ± 0.048	0.29 ± 0.071	0.011 ± 0.001
1000	2.300	0.845 ± 0.037	0.367 ± 0.076	0.011 ± 0.001
1000	2.333	0.881 ± 0.030	0.449 ± 0.079	0.012 ± 0.001
1000	2.367	0.910 ± 0.023	0.538 ± 0.083	0.012 ± 0.001
1000	2.400	0.932 ± 0.017	0.624 ± 0.077	0.012 ± 0.002
1000	2.433	0.948 ± 0.013	0.708 ± 0.071	0.012 ± 0.002
1000	2.467	0.960 ± 0.010	0.783 ± 0.061	0.012 ± 0.002
1000	2.500	0.968 ± 0.008	0.848 ± 0.052	0.013 ± 0.002
1000	2.533	0.974 ± 0.006	0.905 ± 0.045	0.014 ± 0.002

1000	2.567	0.978 ± 0.004	0.949 ± 0.037	0.015 ± 0.002
1000	2.600	0.981 ± 0.003	0.986 ± 0.031	0.016 ± 0.003
1000	2.633	0.984 ± 0.003	1.015 ± 0.026	0.017 ± 0.003
1000	2.667	0.985 ± 0.002	1.038 ± 0.023	0.019 ± 0.003
1000	3.000	0.988 ± 0.002	1.108 ± 0.018	0.029 ± 0.003
1000	3.333	0.987 ± 0.002	1.119 ± 0.019	0.034 ± 0.003
1000	5.000	0.986 ± 0.002	1.130 ± 0.019	0.053 ± 0.003
1000	8.333	0.985 ± 0.003	1.123 ± 0.020	0.071 ± 0.003
1000	16.667	0.985 ± 0.003	1.126 ± 0.020	0.066 ± 0.003
1000	25.000	0.987 ± 0.002	1.137 ± 0.019	0.047 ± 0.003
1000	33.333	0.988 ± 0.002	1.140 ± 0.019	0.038 ± 0.003
10000	0.033	0.001 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
10000	0.067	0.071 ± 0.008	0.002 ± 0.000	0.000 ± 0.000
10000	0.070	0.102 ± 0.014	0.003 ± 0.000	0.000 ± 0.000
10000	0.073	0.139 ± 0.014	0.005 ± 0.000	0.000 ± 0.000
10000	0.077	0.186 ± 0.014	0.007 ± 0.000	0.000 ± 0.000
10000	0.080	0.235 ± 0.016	0.009 ± 0.001	0.000 ± 0.000
10000	0.083	0.286 ± 0.020	0.012 ± 0.001	0.000 ± 0.000
10000	0.087	0.359 ± 0.019	0.016 ± 0.000	0.000 ± 0.000
10000	0.090	0.418 ± 0.023	0.022 ± 0.001	0.000 ± 0.000
10000	0.093	0.475 ± 0.018	0.028 ± 0.001	0.000 ± 0.000
10000	0.097	0.529 ± 0.020	0.035 ± 0.001	0.000 ± 0.000
10000	0.100	0.592 ± 0.020	0.044 ± 0.001	0.000 ± 0.000
10000	0.103	0.636 ± 0.022	0.054 ± 0.001	0.000 ± 0.000
10000	0.107	0.688 ± 0.014	0.067 ± 0.002	0.000 ± 0.000
10000	0.110	0.724 ± 0.019	0.080 ± 0.002	0.000 ± 0.000
10000	0.113	0.756 ± 0.016	0.096 ± 0.003	0.000 ± 0.000
10000	0.117	0.788 ± 0.015	0.113 ± 0.003	0.001 ± 0.000
10000	0.120	0.812 ± 0.014	0.133 ± 0.004	0.001 ± 0.000
10000	0.123	0.835 ± 0.014	0.153 ± 0.003	0.001 ± 0.000
10000	0.127	0.856 ± 0.009	0.176 ± 0.003	0.001 ± 0.000
10000	0.130	0.872 ± 0.009	0.201 ± 0.004	0.001 ± 0.000
10000	0.133	0.887 ± 0.008	0.228 ± 0.005	0.001 ± 0.000
10000	0.137	0.901 ± 0.008	0.257 ± 0.006	0.002 ± 0.000
10000	0.140	0.912 ± 0.007	0.287 ± 0.005	0.002 ± 0.000
10000	0.143	0.922 ± 0.005	0.320 ± 0.006	0.002 ± 0.000
10000	0.147	0.930 ± 0.005	0.354 ± 0.006	0.002 ± 0.000
10000	0.150	0.938 ± 0.005	0.388 ± 0.007	0.003 ± 0.000
10000	0.153	0.946 ± 0.004	0.423 ± 0.008	0.003 ± 0.000
10000	0.157	0.951 ± 0.004	0.459 ± 0.007	0.003 ± 0.000

10000	0.160	0.957 ± 0.003	0.495 ± 0.008	0.004 ± 0.000
10000	0.163	0.961 ± 0.004	0.531 ± 0.008	0.004 ± 0.000
10000	0.167	0.966 ± 0.003	0.567 ± 0.010	0.004 ± 0.000
10000	0.170	0.970 ± 0.003	0.602 ± 0.009	0.005 ± 0.000
10000	0.173	0.972 ± 0.003	0.635 ± 0.010	0.005 ± 0.000
10000	0.177	0.975 ± 0.002	0.668 ± 0.011	0.006 ± 0.000
10000	0.180	0.978 ± 0.002	0.699 ± 0.011	0.006 ± 0.000
10000	0.183	0.980 ± 0.002	0.729 ± 0.011	0.007 ± 0.000
10000	0.187	0.982 ± 0.002	0.757 ± 0.012	0.007 ± 0.000
10000	0.190	0.984 ± 0.002	0.784 ± 0.012	0.007 ± 0.000
10000	0.193	0.985 ± 0.002	0.809 ± 0.012	0.008 ± 0.000
10000	0.197	0.986 ± 0.002	0.833 ± 0.011	0.008 ± 0.000
10000	0.200	0.987 ± 0.002	0.853 ± 0.012	0.009 ± 0.000
10000	1.333	0.996 ± 0.001	1.120 ± 0.004	0.023 ± 0.001
10000	1.667	0.996 ± 0.001	1.118 ± 0.004	0.021 ± 0.001
10000	2.000	0.996 ± 0.001	1.117 ± 0.004	0.019 ± 0.001
10000	2.333	0.996 ± 0.001	1.115 ± 0.004	0.018 ± 0.001
10000	2.667	0.996 ± 0.001	1.115 ± 0.004	0.017 ± 0.001
10000	3.000	0.996 ± 0.001	1.114 ± 0.004	0.017 ± 0.001
10000	3.333	0.996 ± 0.001	1.114 ± 0.004	0.016 ± 0.001
10000	5.000	0.997 ± 0.001	1.113 ± 0.004	0.013 ± 0.001
10000	8.333	0.997 ± 0.001	1.113 ± 0.004	0.010 ± 0.001
10000	16.667	0.997 ± 0.001	1.112 ± 0.004	0.007 ± 0.001
10000	25.000	0.997 ± 0.001	1.110 ± 0.004	0.005 ± 0.001
10000	33.333	0.998 ± 0.001	1.107 ± 0.004	0.004 ± 0.001

Table 4: *Mean ± SD r^2, slopes and intercepts per SNP array-like subsampling for the small population. Means are among subsampling and simulation replicates. Exact Mean r^2, slopes and intercepts per simulation replicate can be found in Table S5. Rows are ordered by array size.*

N Population	**ARRAY**	r^2	**Slope**	**Intercept**
1'000	50k	0.989 ± 0.002	1.036 ± 0.015	0.012 ± 0.002
1'000	700k	0.985 ± 0.003	1.127 ± 0.014	0.056 ± 0.003

Supplementary Figures

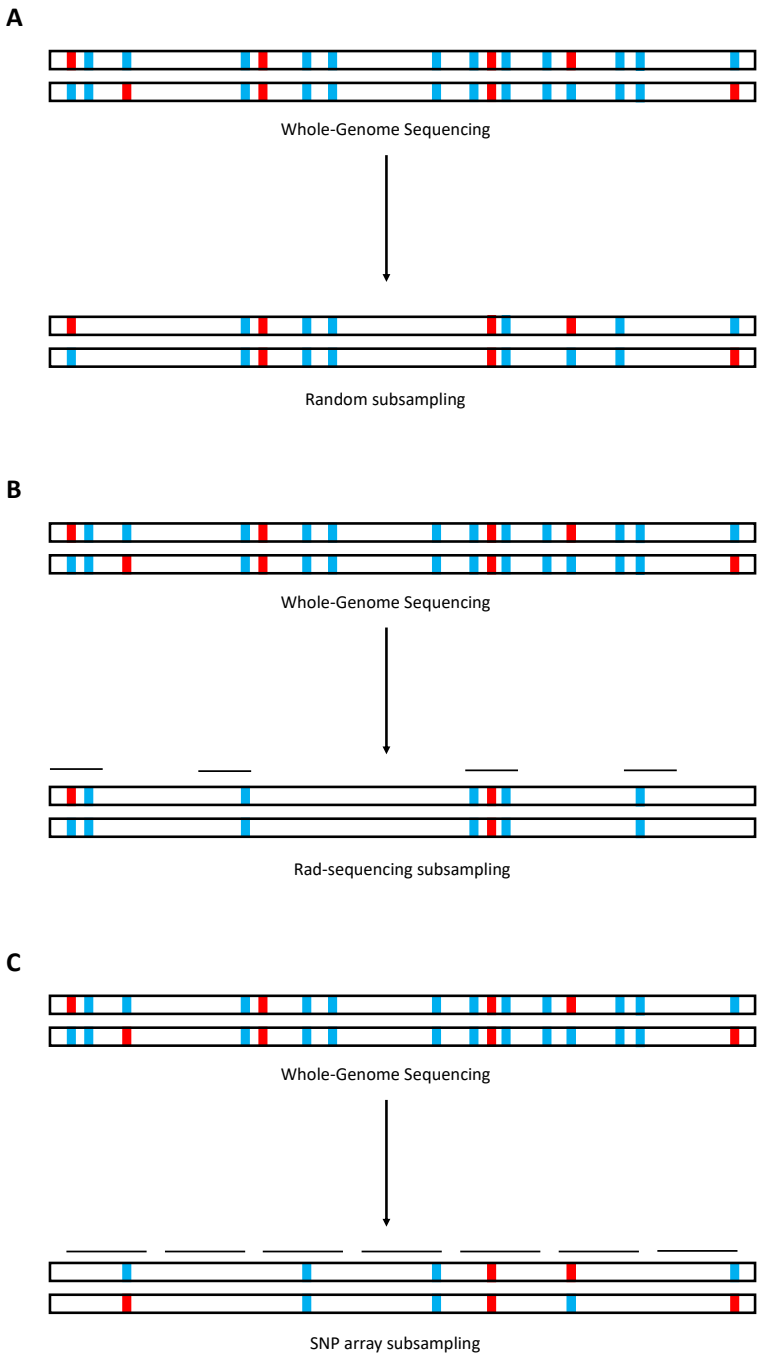

Figure S1*: Scheme of three of the four subsampling techniques for a small genomic segment. Each square is a SNP, blue SNPS are the reference allele and red SNPs are the derived allele.* ***A:*** *Random subsampling, SNPs are completely randomly selected****. B:*** *Rad-sequencing, 500bp windows (represented by the black lines above the genome) are randomly assigned in the genome and each SNP within these windows gets subsampled.* ***C:*** *SNP array, windows (represented by the black lines above the genome) are equally spaced in the genome. If there is SNPs in the window, one is randomly selected.*

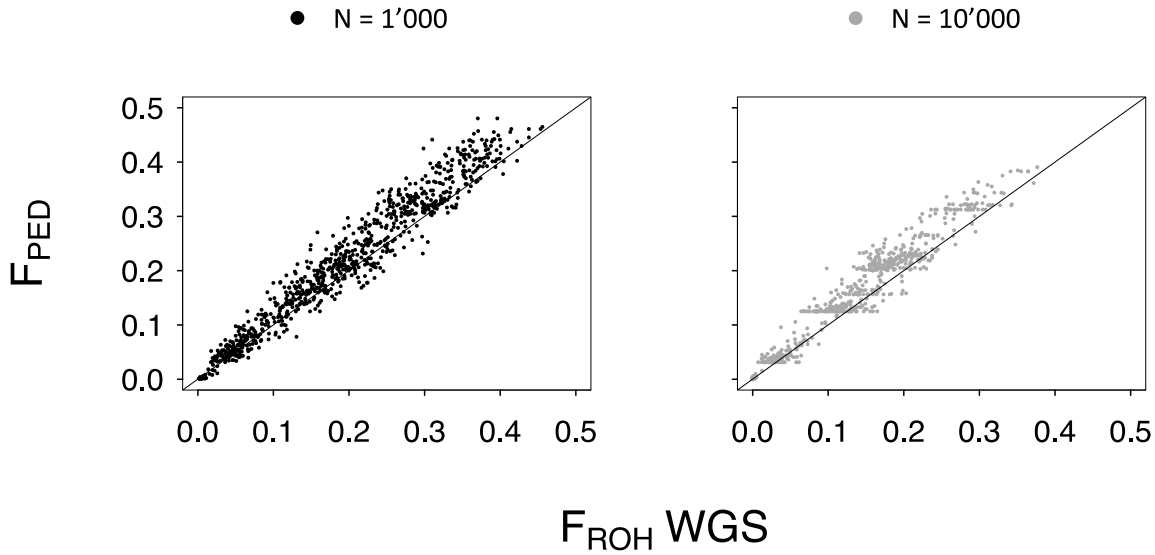

Figure S2: *The relationship between F_{PED} on the y axis and F_{ROH} calculated with all the data (WGS-like) on the x axis for both populations (small is in back on the left and large is in grey on the right). The black lines represent the perfect correlation between both. Each point represents one individual. All simulation replicates are represented in these graphs. r^2: 0.975, slope: 1.077 and intercept: 0.007 for the small population. r^2: 0.975, slope: 1.079 and intercept: 0.009 for the large population.*

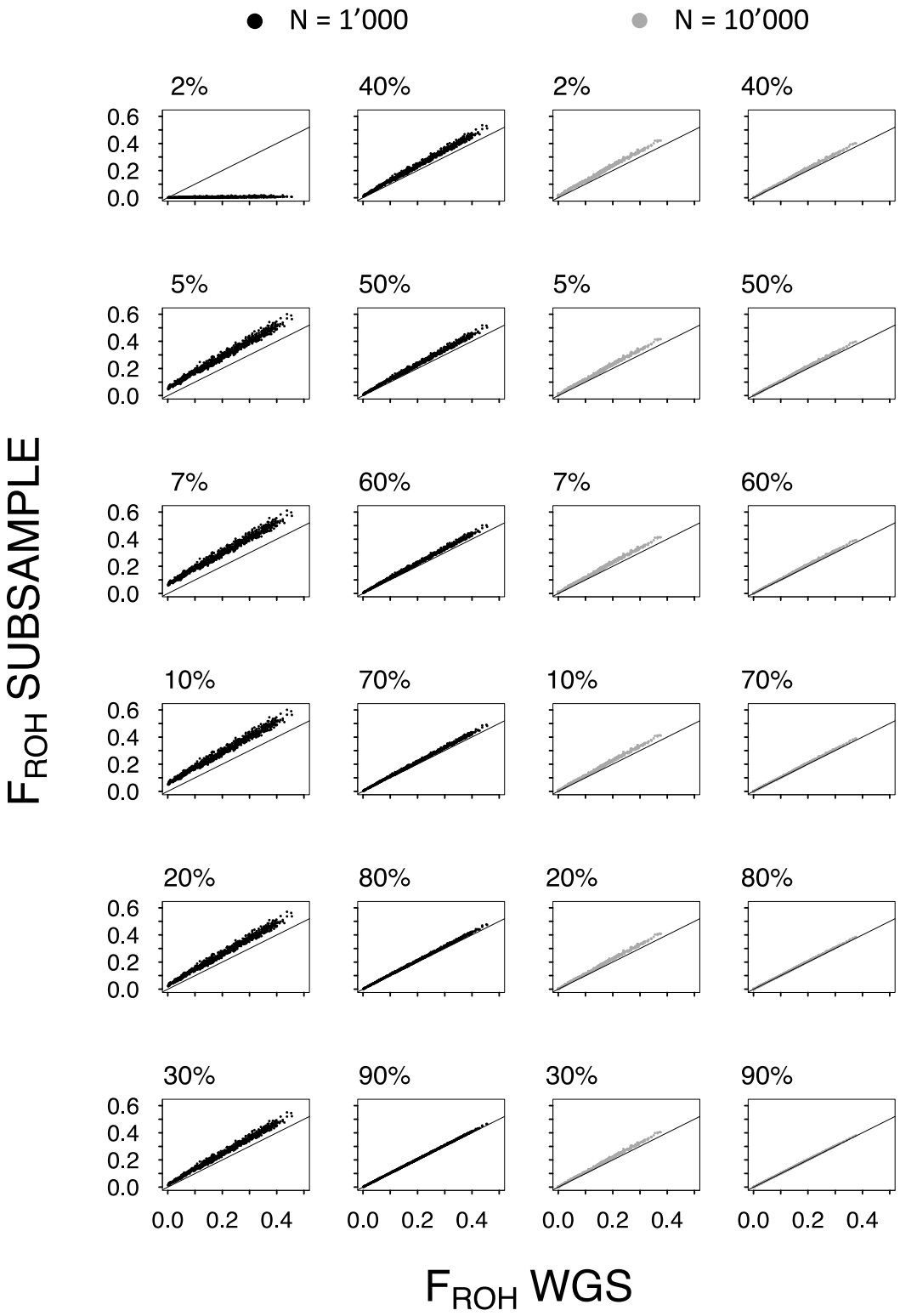

Figure S3: *The relationship between F_{ROH} calculated with a subsample of the SNPs (indicated in percentages above plots) in the y axis and F_{ROH} calculated with all the SNPs in the x axis for both populations (small is in back on the two left columns and large is in grey on the two right columns). The black lines represent the perfect correlation between both. Each point represents one individual. All simulation replicates are represented in these graphs. F_{ROH} subsampled is represented as the mean among subsampling replicates. r^2, slopes and intercepts are reported in Table S2.*

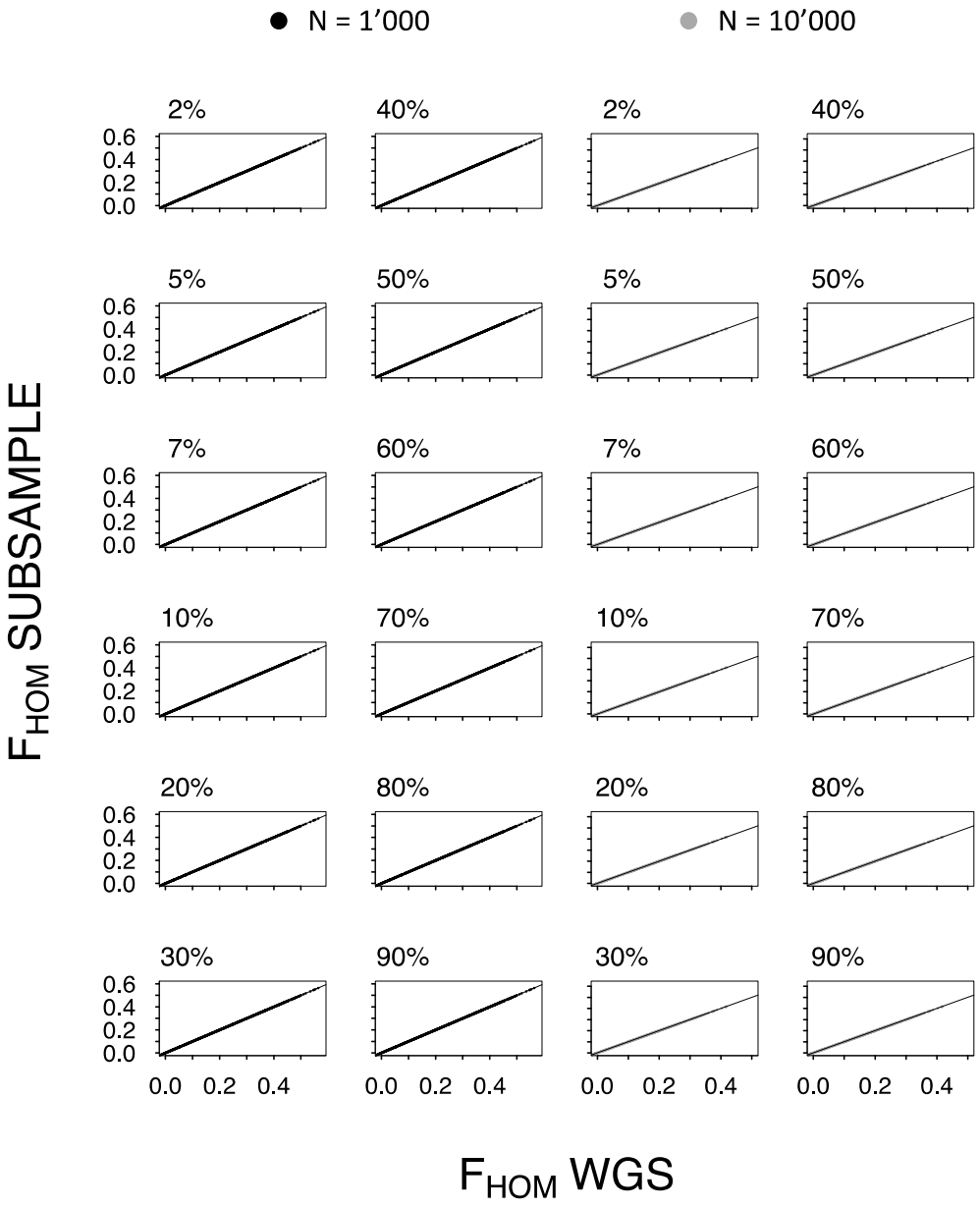

Figure S4: *The relationship between F_{HOM} calculated with a subsample of the SNPs (indicated in percentages above plots) in the y axis and F_{HOM} calculated with all the SNPs in the x axis for both populations (small is in back on the two left columns and large is in grey on the two right columns). The black lines represent the perfect correlation between both. Each point represents one individual. All simulation replicates are represented in these graphs. F_{HOM} subsampled is represented as the mean among subsampling replicates.*

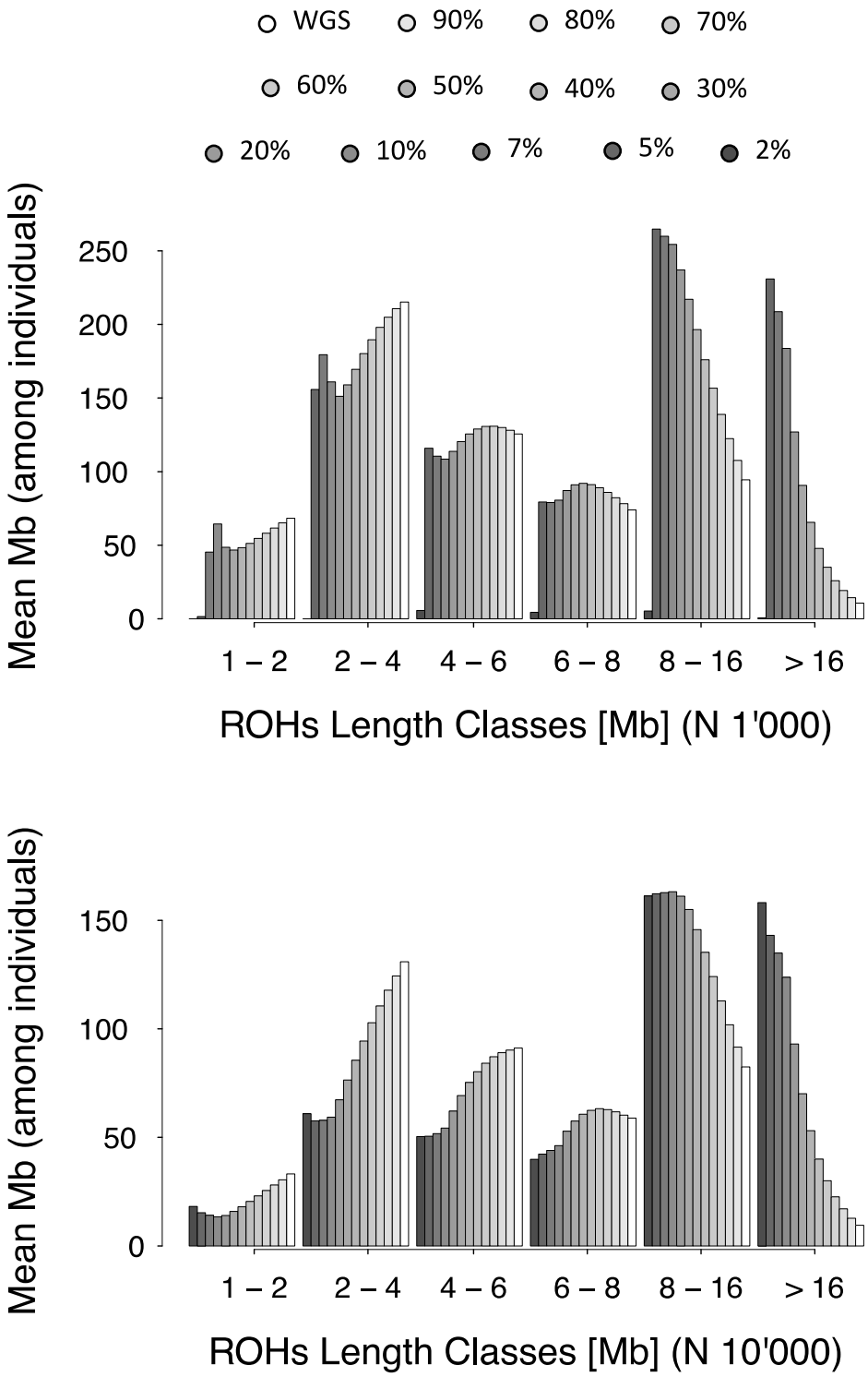

Figure S5: *ROHs distributions. The mean individual sum of lengths of ROHs on the y axis per ROHs length classes on the x axis. Small population is above and large population is below. Individual means are among subsampling and simulations replicates. Legend for both plots on top of the first plot.*

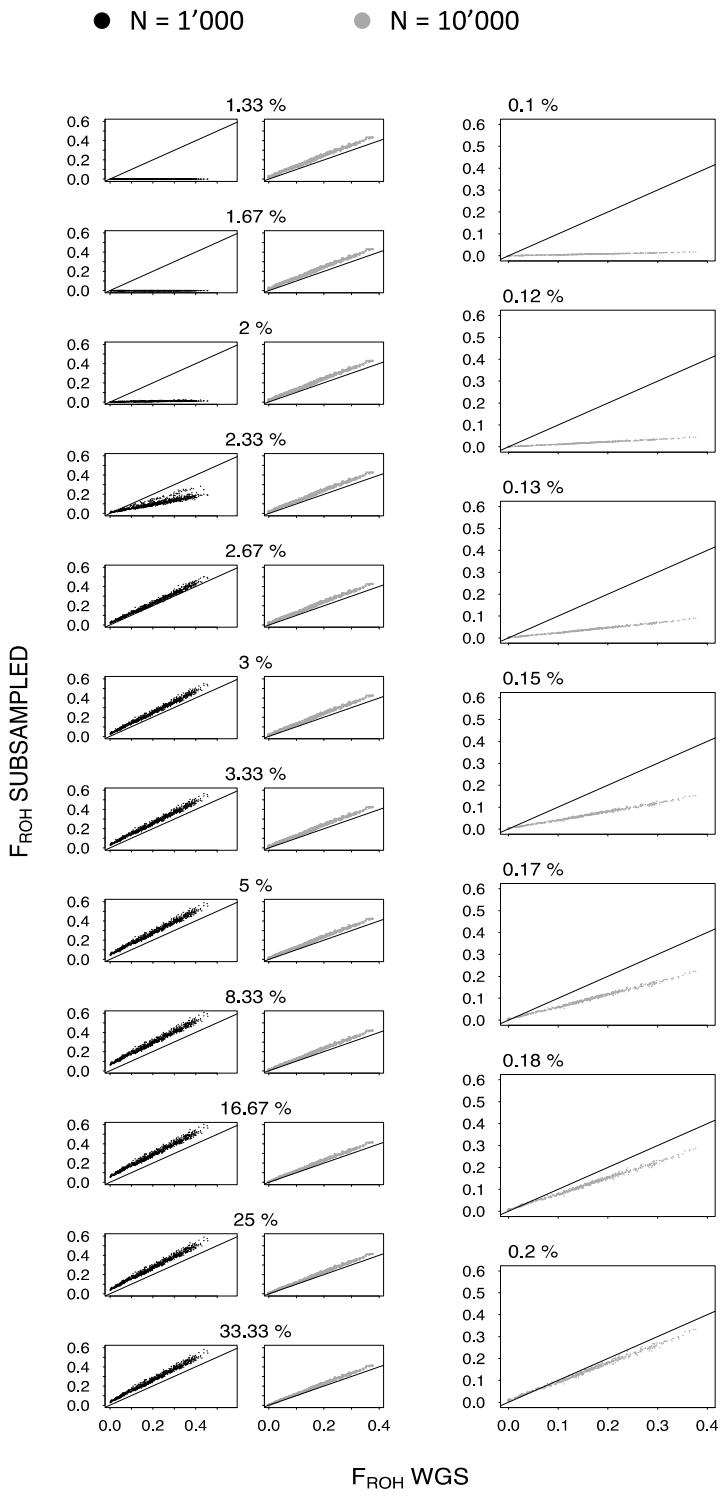

Figure S6: *The relationship between F_{ROH} estimated with a subsample of the genome (indicated in percentages above plots) mimicking RAD-sequencing on the y axis and F_{ROH} calculated with all the data (WGS-like) on the x axis for both populations (small is in back on the left and large is in grey on the right). The percentages represent the proportion of genome sequenced. The black lines represent the perfect correlation between both. Each point represents one individual. All simulation replicates are represented in these graphs. F_{ROH} RAD-sequencing is represented as the mean among subsampling replicates. r^2, slopes and intercepts are reported in Table S4.*

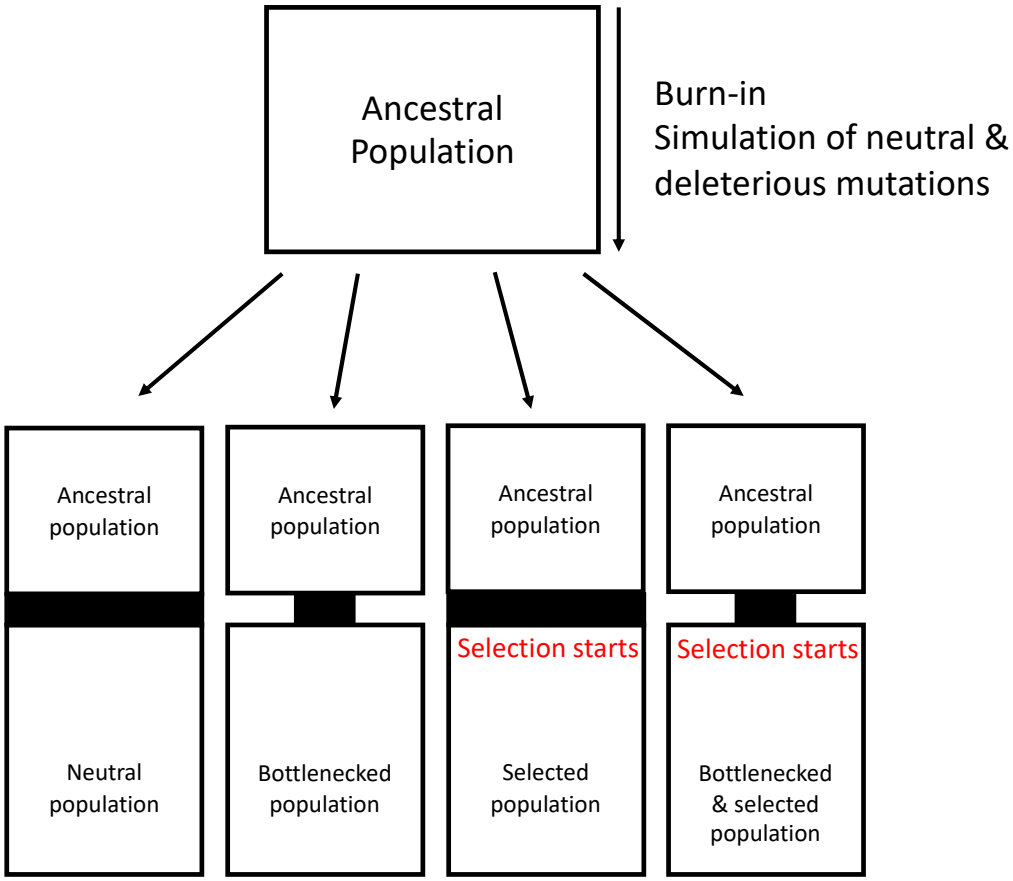

Figure S7: *Scheme of Simulations. One burn in will be performed and then be used to construct the four populations. No change for the first population, a bottleneck for the second population, apparition of selection for the third population and a bottleneck and the apparition of selection for the fourth population.*

Supplementary tables

Table S1: *Number of individuals and SNPs per simulation replicate. Rows are ordered by population size first and simulation replicate second.*

N Population	Simulation ID	N individuals	N SNPs
1000	Sim1668789343933	85	2’532’168
1000	Sim1668809343933	92	2’565’701
1000	Sim1668829343933	97	2’541’966
1000	Sim1668839343936	90	2’549’833
1000	Sim1668849343939	84	2’537’280
1000	Sim1669039565968	84	2’585’239
1000	Sim1669059566969	81	2’558’508
1000	Sim1669079572071	83	2’552’727
1000	Sim1669099572564	84	2’528’199
1000	Sim1669119579366	84	2’642’356
10’000	Sim13281	69	34’795’044
10’000	Sim16055	69	35’491’494
10’000	Sim16887	66	35’453’865
10’000	Sim21523	71	35’393’905
10’000	Sim32005	64	35’506’361
10’000	Sim5258	72	35’422’842
10’000	Sim5942	68	35’540’937
10’000	Sim6008	64	35’516’626
10’000	Sim8007	61	35’408’613
10’000	Sim8856	60	35’564’205